ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANABİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

AKADEMİK PERSONEL VE LİSANSÜSTÜ EĞİTİMİ GİRİŞ SINAVININ BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TEST OLARAK UYGULANABİLİRLİĞİNİN İNCELENMESİ

DOKTORA TEZİ

TANSU ALAN

ANKARA
OCAK, 2026


ANKARA ÜNİVERSİTESİ
1946

ANKARA ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

EĞİTİM BİLİMLERİ ANABİLİM DALI
EĞİTİMDE ÖLÇME VE DEĞERLENDİRME PROGRAMI

AKADEMİK PERSONEL VE LİSANSÜSTÜ EĞİTİMİ GİRİŞ SINAVININ BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TEST OLARAK UYGULANABİLİRLİĞİNİN İNCELENMESİ

DOKTORA TEZİ

TANSU ALAN

DANIŞMAN: PROF. DR. SEHER YALÇIN

ANKARA
OCAK, 2026

Ankara Üniversitesi Eğitim Bilimleri Enstitüsü Müdürlüğüne,

TANSU ALAN adlı öğrencinin hazırladığı "Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavının Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi" başlıklı bu çalışma Eğitim Bilimleri Anabilim Dalı / Eğitimde Ölçme ve Değerlendirme Programı'nda jüri üyelerince oy birliği ile **Doktora Tezi** olarak kabul edilmiştir.

	Jüri Üyeleri	İmza
Başkan	Doç. Dr. Eren Can AYBEK	……..
Üye	Prof. Dr. Seher YALÇIN (Danışman)	……..
Üye	Prof. Dr. Kaan Zülfikar DENİZ	……..
Üye	Prof. Dr. Celal Deha Doğan	……..
Üye	Doç. Dr. İlker Kalender	……..

ONAY

Bu tez Ankara Üniversitesi Lisansüstü Eğitim Öğretim Yönetmeliği'nin ilgili maddeleri uyarınca, jüri üyeleri tarafından …/…/20... tarihinde, Enstitü Yönetim Kurulu tarafından ise …/…/20... tarihinde kabul edilmiştir.

……

Prof. Dr. Mehmet İkbal YETİŞİR

Eğitim Bilimleri Enstitüsü Müdürü

ETİK İLKELERE UYGUNLUK BİLDİRİMİ

Tez içindeki bütün bilgileri akademik yazım kurallarına uygun biçimde raporlaştırdığımı ve bunları etik ilkelere (atıfta bulunulan tüm yapıtlara kaynaklarda yer verilmesi, tezde kullanılan bilgi ve belgelere resmi yollarla ulaşılması ve bunların aslı bozulmadan kullanılması vb.) uygun olarak elde ettiğimi ve sunduğumu bildiririm.

Tansu ALAN

ÖZET

AKADEMİK PERSONEL VE LİSANSÜSTÜ EĞİTİMİ GİRİŞ SINAVININ BİLGİSAYAR ORTAMINDA BİREYE UYARLANMIŞ TEST OLARAK UYGULANABİLİRLİĞİNİN İNCELENMESİ

ALAN, Tansu
Doktora Tezi, Eğitimde Ölçme ve Değerlendirme Anabilim Dalı
Tez Danışmanı: Prof. Dr. Seher YALÇIN
Ocak, 2026, xvi + 151

Bu araştırmanın amacı, kâğıt-kalem (KK) formatında uygulanan Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES)'nın Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) olarak uygulanabilirliğini incelemektir. Bu doğrultuda, Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM)'nden temin edilen verilerle post-hoc simülasyonları gerçekleştirilmiştir. Simülasyonlarda madde kullanım sıklığı ve kapsam dengelemesi tüm koşullarda sabit tutularak, yetenek kestirim yöntemi olarak (Beklenen Sonsal Dağılım-BSD ve Maksimum Olabilirlik-MO), test sonlandırma kuralı olarak (standart hata SH < 0.30, < 0.40, <0.50 ve sabit madde sayısı MS = 15, 20) ve madde seçim yöntemi olarak (Maksimum Fisher Bilgisi, Kullback-Leibler Bilgisi, Ağırlıklandırılmış Olabilirlik Bilgisi, Verimlilik Dengeli Bilgi) kullanılmıştır. Simülasyon sonuçlarına göre, alt testlerin BOBUT uygulaması için ölçme hassasiyeti açısından uygun koşulların; yetenek kestirim yöntemi olarak BSD, madde seçim yöntemi olarak Maksimum Fisher Bilgisi ve test sonlandırma kuralı olarak MS = 20 olduğu belirlenmiştir. Post-hoc simülasyon sonuçlarına dayanarak, ALES'in canlı BOBUT uygulaması geliştirilmiş ve 101 katılımcıya hem ALES KK hem de ALES BOBUT uygulanmıştır.

Araştırma bulgularına göre, katılımcıların ALES sayısal alt test canlı BOBUT uygulamasından elde ettikleri yetenek düzeyleri ile ALES KK uygulamasından elde ettikleri yetenek düzeyleri arasında anlamlı bir fark bulunmamış ve yetenek düzeyleri arasında .75 düzeyinde pozitif bir ilişki saptanmıştır. Buna karşın, sözel alt testte katılımcıların canlı BOBUT ve KK uygulamalarından elde ettikleri yetenek düzeyleri arasında anlamlı bir fark bulunmuştur. Ayrıca, yetenek düzeyleri arasında .59 düzeyinde

pozitif bir ilişki saptanmıştır. Kapsam dengelemesi açısından incelendiğinde, her iki alt testte de canlı BOBUT uygulamasında içerik dengelemesinin sağlandığı görülmüştür.

Araştırma sonuçları, ALES'in sayısal alt testinde BOBUT'un, KK ile benzer ölçme gücüne sahip olduğunu ve daha az maddeyle daha hassas kestirimler sağlayabildiğini göstermektedir. Sözel alt testin BOBUT olarak uygulanabilirliğini artırmak için ise, okuma metinlerinin uzunluğunun ve yoğunluğunun optimize edilmesi, katılımcıların istediği alt testten başlama esnekliğinin sunulması ve mevcut BOBUT yaklaşımının, modül bazlı Çok Aşamalı Bireye Uyarlanmış Test (Multistage Adaptive Testing, MST) modeliyle karşılaştırılması önerilmektedir.

Anahtar Sözcükler: Bilgisayar Ortamında Bireye Uyarlanmış Test, Madde Tepki Kuramı, Ölçme ve Değerlendirme, ALES, BOBUT

ABSTRACT

AN INVESTIGATION OF THE APPLICABILITY OF THE ACADEMIC PERSONNEL AND GRADUATE EDUCATION ENTRANCE EXAMINATION AS A COMPUTERIZED ADAPTIVE TEST

ALAN, Tansu

Doctoral Dissertation, Department of Measurement and Evaluation in Education

Thesis Advisor: Prof. Dr. Seher YALÇIN

January, 2026, xvi + 154 Pages

The purpose of this study is to examine the applicability of the Academic Personnel and Graduate Education Entrance Examination (ALES), which is administered in a paper-and-pencil (PP) format, as a Computerized Adaptive Test (CAT) version. For this purpose, post-hoc simulations were conducted using data obtained from the Measurement, Selection and Placement Center (ÖSYM). In the simulations, item exposure control and content balancing were held constant across all conditions, while ability estimation methods (Expected a Posteriori [EAP] and Maximum Likelihood [ML]), test termination rules (standard error thresholds of SE < 0.30, < 0.40, < 0.50, and fixed test lengths of 15 and 20 items), and item selection methods (Maximum Fisher Information, Kullback–Leibler Information, Likelihood-Weighted Information, and Efficiency-Balanced Information) were systematically varied.

Simulation results indicated that the optimal conditions for CAT administration in terms of measurement precision were the use of the EAP ability estimation method, Maximum Fisher Information item selection, and a fixed test length of 20 items. Based on these post-hoc simulation results, a live CAT version of ALES was developed and administered to 101 participants alongside the traditional PP ALES.

The findings showed that there was no statistically significant difference between ability estimates obtained from the live CAT and the PP administration for the quantitative subtest, and a strong positive correlation (r = .75) was observed between the two administrations. In contrast, a statistically significant difference was found between CAT and PP ability estimates for the verbal subtest, favoring the PP format, with a

moderate positive correlation (r = .59) between the two sets of ability estimates. Examination of content balancing revealed that content constraints were successfully met in the live CAT administration for both subtests.

Overall, the results demonstrate that the CAT version of the quantitative subtest of ALES provides measurement precision comparable to the PP format while achieving more accurate ability estimation with fewer items. To enhance the applicability of CAT for the verbal subtest, it is recommended to optimize the length and density of reading passages, allow examinees flexibility in choosing the starting subtest, and compare the current CAT approach with a module-based Multistage Adaptive Testing (MST) model.

Keywords: Computerized Adaptive Testing, Item Response Theory, Measurement and Evaluation, ALES, CAT

ÖNSÖZ

Bu tez, Türkiye'deki sınavların büyük bir bölümünü yürüten önemli bir test merkezi olan Ölçme, Seçme ve Yerleştirme Merkezi (ÖSYM) tarafından uygulanan Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'nın (ALES), Madde Tepki Kuramı çerçevesinde Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) formatında uygulanabilirliğini incelemektedir. Klasik Test Kuramı ve geleneksel ölçme değerlendirme süreçlerinden farklı olarak, BOBUT yaklaşımı daha kısa test formları ile daha adil, daha objektif, daha güvenilir ve geçerli ölçme sonuçları sunma potansiyeline sahiptir. Dünyada dijitalleşmenin yaygınlaşması, teknolojiye erişimin kolaylaşması ve önde gelen uluslararası test merkezlerinin (Educational Testing Service- ETS, Graduate Management Admission Council-GMAC gibi) sınavlarını BOBUT olarak uygulamaları göz önünde bulundurulduğunda, Türkiye'de gerçekleştirilen sınavların da bu formatta uygulanmasının önemli faydalar sağlayacağı öngörülmektedir. Bu bağlamda, tez kapsamında elde edilen bulguların ulusal sınav uygulamalarına katkı sunması beklenilmektedir.

Doktora eğitimime pandemi döneminde, uzaktan eğitim koşullarında başladım. Bu sürecin en başında tanışma vesilesiyle attığı içten bir e-posta ile gönlümü kazanan, doktora yolculuğum boyunca akademik ve etik rehberliğinin yanı sıra insani yönüyle de bir "başucu kitabı" olan hem akademik hem de kişisel duruşuna hayranlık duyduğum çok kıymetli danışmanım Prof. Dr. Seher Yalçın'a en derin teşekkürlerimi sunuyorum. Yeri geldiğinde bir dost, yeri geldiğinde yol gösterici olan ve bu sürecin her aşamasında "iyi ki" dedirten, varlığıyla "Seher Hocam varsa bu tez biter" güveni bana her daim hissettirdi. Hayatım boyunca ne kadar teşekkür etsem az kalır.

Doktora eğitimim süresince, bilgi ve deneyimleriyle akademik kimliğimin şekillenmesinde büyük katkıları olan, aynı zamanda TİK üyeliğiyle rehberliğini esirgemeyen değerli Prof. Dr. Kaan Zülfikar Deniz hocama,

Tezimin en zor ve çıkmaza girdiği anlarda, yaptığı yapıcı müdahalelerle çalışmamın tamamlanmasında destek sağlayan, çok kıymetli TİK üyesi Doç. Dr. Eren Can Aybek hocama,

Görüş ve önerileriyle bu araştırmaya katkı sağlayan çok kıymetli juri üyesi Doç. Dr. İlker Kalender hocama,

Akademik hayatın temel taşı olan etik değerleri, çalışma disiplinini, bilgiye ulaşma ve bilgiyi üretme yollarını öğreten; Ankara Üniversitesi'nde ders alma ayrıcalığına eriştiğim ve her biri benim için rol model, dünyalar iyisi dünyalar tatlısı olan Prof. Dr. Celal Deha Doğan, Prof. Dr. Hamide Deniz Gülleroğlu, Prof. Dr. Ergül Demir hocalarıma,

TÜBİTAK 1001 projesinde bursiyer olduğum, uzun toplantılar boyunca hem çalışıp hem eğlendiğimiz, enerjisi yüksek ve sevecen Doç. Dr. Selen Demirtaş Zorbaz hocama,

Ölçme ve değerlendirme alanındaki ilk öğrenmelerimin sahibi olan akademik ve psikolojik olarak yıllardır desteğini aldığım lisanstan çok değerli Doç. Dr. Serbay Duran hocama,

Doktora tezimi yazma sürecinde, çalışmalarımı İspanya'nın kuzeyinde yer alan León Üniversitesi'nde sürdürme olanağı tanıyan; misafirperverliği, yardımseverliği, mentörlüğü ve pozitif enerjisiyle destek sağlayan Prof. Dr. Ana Rosa Arias Gago hocama,

Motivasyon, destek ve samimiyetleriyle bana güç veren değerli arkadaşlarıma,

Bu çalışmada kullanılan verilerin paylaşımı konusundaki hassasiyetleri ve gösterdikleri yakın ilgi nedeniyle ÖSYM'ye, ÖSYM nezdinde; ÖSYM ARGE ve ÖSYM Bilgi İşlem birimlerine,

Akademik kariyer yolculuğumda ve tüm eğitim hayatımda her zaman yanımda olan, beni destekleyen, cesaretlendiren ve fedakârlıklarını esirgemeyen; yüksek lisanstayken doktorayı, doktoradayken doçentliği sorarak hep daha ileriye yönlendiren, doğup büyüdüğüm coğrafyada eğitime verdiği önem ve adı gibi aydın kişiliğiyle bana örnek olan emektar canım babam Aydın ALAN'a ve aileme; ayrıca bilgisayarı açtığım anda yanıma gelip saatlerce izleyen, tüm stresimi alan, varlığı ve sevimli tavırlarıyla enerji veren tatlı kedimiz Keje'ye…

Tüm kalbimle teşekkürlerimi sunuyorum. İyi ki hayat akışımın önemli dönemlerinde var oldunuz.

Bu tezi güzel günlerin geleceğini umutla ve sabırla bekleyen tüm çocuk işçilere armağan ediyorum.

İÇİNDEKİLER

Sayfa

TABLOLAR DİZİNİ

Tablo **Sayfa**

ŞEKİLLER DİZİNİ

Şekil **Sayfa**

BÖLÜM 1

GİRİŞ

Bu bölümde araştırmanın problem durumu açıklanmış, devamında araştırmanın amacına, önemine, sınırlıklarına ve tanımlara yer verilmiştir.

Problem Durumu

Sabit uzunluktaki kâğıt-kalemle yapılan geleneksel testler, 1900'lerin başından bu yana eğitim değerlendirmesinin temelini oluşturmuştur. Eğitimde yapılan testlerin büyük bir çoğunluğu bu kategoriye girerken, bireysel olarak uygulanan zekâ ve teşhis testleri istisna olarak kabul edilir. 1940'lara kadar, bu tür testlerin neredeyse tamamı klasik test kuramı (KTK) kullanılarak değerlendirilmektedir (Gulliksen, 1950).

Bilgisayar teknolojisinin gelişimi toplumun birçok yönünü etkilediği gibi, testler üzerinde de etkili olmuştur. Bilgisayarların kullanılabilirliği ve psikometri alanında çalışan araştırmacıların (Lord, 1980; Lord ve Novick, 1968; Weiss, 1982; Wright ve Stone, 1979) Madde Tepki Kuramı (MTK) üzerindeki çalışmaları testlerin iyileştirilmesine yönelik giderek daha karmaşık araştırmaları ve gelişmeleri teşvik etmiştir.

MTK'nın ortaya çıkışı ve ilerlemesi ile yeni bilişim teknolojilerinin hızlı kullanımı eğitim testlerinin tasarımı, sunumu ve uygulanmasını tamamen değiştirmiştir. Bu değişimlerin en önemli örneklerinden biri bilgisayar ortamında bireye uyarlanmış test [Computerized Adaptive Test -CAT (Bilgisayar Ortamında Bireye Uyarlanmış Test BOBUT)] sistemidir (Han, 2018). Kağıt-kalem veya bilgisayar destekli olan doğrusal testlerde tüm katılımcılara belirli bir test formundaki tüm maddeler sunulmaktadır. Aynı beceriyi ölçen bir testin zaman içinde tekrar kullanılması için yeni maddeler yazılmakta bu durum ise zaman ve maliyet gerektiren bir çaba olabilmektedir. Buna karşın BOBUT uygulamaları farklı bir yaklaşım benimsemektedir. BOBUT'ta maddeler test esnasında bireyin yetenek düzeyine uygun olacak şekilde seçilir. Her katılımcının maddelere verdiği yanıtlar test sırasında kaydedilir ve katılımcının yetenek düzeyi düzenli olarak

güncellenen bir tahmini korur. Sonuç olarak katılımcılara aynı testin bireyselleştirilmiş versiyonları sunulur (Wainer, 2000).

BOBUT uygulamalarının ilk örnekleri 1900'lerin başlarında Binet'in zekâ testlerine dayanmasına rağmen, ABD'nin test kurumu olan Eğitim Test Servisi (Educational Testing Service-ETS)'nde ciddi araştırmalar yapan Fred Lord'un katkılarıyla bu alandaki çalışmalar şekillenmeye başlanmıştır. Lord'un amacı düşük ve yüksek yetenek düzeyine sahip bireylerin yeteneğini hassas bir şekilde ölçen bir test geliştirmektir. Lord'un teorik olarak tanımladığı bu test, ölçme hassasiyetini azaltmadan test uzunluğunu kısaltacaktır. Çünkü test bireyin yetenek düzeyi hakkında en fazla bilgiyi içeren maddelerden oluşacaktır. Bilgisayar teknolojisinin gelişmesiyle birlikte bu teorik BOBUT gerçeklik kazanmış ve BOBUT uygulamaları yaygınlaşmıştır (Hogan, 2014).

Son kırk yılda ise kağıt-kalem testlerine karşı başarılı bir alternatif olarak BOBUT kullanımında bir artış yaşanmıştır (Luecht ve Sireci, 2011). 1993'te ETS ilk geniş ölçekli BOBUT uygulamasını lisansüstü eğitimde öğrenci seçim sınavları olan GRE-CAT (Lisansüstü Eğitim Giriş Sınavı-Graduate Record Examination-CAT) ve devamında 1997'de ETS işletme ve yönetim alanında yüksek lisans programlarına kayıt için GMAT (İşletme ve Yönetim Programları Lisansüstü Eğitim Giriş Sınavı-Graduate Management Admission Test) BOBUT versiyonunu geliştirmiştir. Hemen ardından Temmuz 1998'de ise Yabancı Dil Olarak İngilizce Sınavı (Test of English as a Foreign Language-TOEFL), BOBUT olarak uygulanmaya başlanmış ve bu uygulama 2001 yılı itibariyle kağıt-kalem testlerinin yerini almaya başlamıştır (Slater, 2001). Amerika'da BOBUT uygulamaları psikolojik ve işe alım testlerinin birçok türünde de kullanılmaktadır. Kansas, Oregon, Teksas ve Virginia gibi birçok eyalet artık sınıf sonu, ders sonu ve lise mezuniyeti için BOBUT seçenekleri sunmaktadır (Luecht ve Sireci, 2011). Dünya geneline bakıldığında; Japonya'da J-CAT (Japenese Computerized Adaptive Test), Amerika ve Latin Amerika'da The CAT (The CAT of Written English for Spanish Speakers), Amerika'da CATE (Computerized Adaptive Test of English), Amerika ve Latin Amerika'da STAR Reading, STAR Math, STAR Earliey Literacy ve STAR Reading Spanish uygulanan BOBUT örnekleridir (Kezer, 2020).

BOBUT uygulamaları gerçekleştirilirken bireyin yetenek düzeyini tahmin etmek için madde havuzundan belirli stratejiler altında madde seçilir. Bu stratejiler sıklıkla MTK kullanılarak uygulanır (Weiss ve Kingsburry, 1984). MTK bireylerin test performansı ile performansın altında yatan gizil özellikler arasındaki ilişkiyi tanımlamaya çalışan matematiksel bir modeldir (Hambleton ve Swaminathan, 1985). MTK'da bireyin yetenek

düzeyi ile bir maddeye doğru cevap verme olasılığı arasındaki ilişki, madde karakteristik eğrisi (MKE) adı verilen bir eğri yardımıyla hesaplanır (Reckase, 2009). MTK, BOBUT teknolojisi ilk geliştirilmeye başlandığında tanımlanan birçok soruna cevap sağlamıştır. Bu sorunlara MTK; maddeleri tek bir ölçek üzerine yerleştirme, birey ve madde özelliklerini ortak bir ölçekte bir araya getirme, madde seçimi için güçlü algoritmalar sağlama ve bilgi fonksiyonlarının kullanımıyla test tasarımına izin verme gibi tatmin edici çözümler sunmuştur. BOBUT, şu anda MTK'nın en başarılı uygulamalarından biri olarak düşünülebilir. MTK kullanılarak geliştirilen BOBUT'un tüm sorunlarını çözme imkânı bulunmasa da çözülen sorunlar, BOBUT'u ileri bir uygulama aşamasına taşımıştır. (Kingsbury ve Houser, 2005).

BOBUT uygulamaları, bireyin maddeye verdiği cevaba dayalı olarak madde havuzundan yeni madde seçilen bir test sürecidir. BOBUT'ta bireye bir madde uygulanır ve birey maddeyi yanıtlar, ardından verilen yanıt puanlanır, bireyin yetenek düzeyi kestirilir ve kestirilen yetenek düzeyine uygun yeni bir madde seçilir (Weiss ve Kingsburry, 1984). BOBUT uygulamalarında bireylere, kendi yetenek düzeylerine en uygun maddeler sunulmaktadır. Uygulanacak madde, bireyin önceki maddelere verdiği yanıtlardan hareketle hesaplanan geçici yetenek kestirimine göre seçilir. Bu süreçte, bireyin mevcut yetenek kestirimine en fazla bilgi sağlayan madde belirlenerek uygulanır. Bu yönüyle BOBUT, geleneksel kâğıt-kalem testlerine kıyasla daha kısa testlerle hedeflenen ölçme kesinliğine ulaşmayı mümkün kılmaktadır (Weiss, 1982).

BOBUT uygulamaları beş bileşenden oluşur. İlk bileşeni kalibre edilmiş bir madde havuzudur ve ölçülmek istenen yapının içeriğine bağlı olarak geliştirilir. Diğer dört bileşen olan test başlatma, madde seçim yöntemleri, yetenek kestirim yöntemleri ve sonlandırma kuralı ise içerikten ziyade BOBUT'ta uygulanan testin güvenirlik ve geçerliğine ait özellikleridir (Thompson ve Weiss, 2011).

BOBUT uygulamalarında tüm yetenek düzeylerine hitap eden madde parametreleri ve MTK'ya göre kestirilmiş geniş bir madde havuzuna ihtiyaç vardır (Green vd., 1984). Madde havuzunun ne kadar büyük olduğu testin kullanım amacına göre karar verildiği gibi Monte Carlo gibi simülasyonlarla da karar verilebilir (Thompson ve Weiss, 2011). Madde havuzunun BOBUT uygulamasına katılacak her bireyin yetenek düzeyine uygun maddelerden oluşması önemlidir. Örneğin çok zor maddelerin olduğu bir madde havuzu yetenek düzeyi düşük bireyler için ayırt edici değilken aynı şekilde kolay maddelerin olduğu bir madde havuzu da yetenek düzeyi yüksek bireyler için ayırt edici değildir. Bunun için madde havuzunun geniş bir yetenek ranjına hitap etmesi gerekir.

BOBUT uygulamalarında başlatma kuralı farklı yöntemlerle yapılabilmektedir. Bu uygulamalarda çoğunlukla bireylerin başlangıç teta düzeyleri sıfıra eşitlendiği ($\theta = 0$) gibi bireylerin önsel dağılımına ait bilgiler de başlatma kuralı olarak kullanılmaktadır. Başlatma kuralı $\theta = 0$ alındığında bireylere uygulanacak ilk maddenin havuzdaki aynı madde olma ihtimali yüksektir bunun için θ'nın "-0.5 ve 0.5" arasında seçilmesi aynı maddenin uygulanma olasılığını düşürecektir. Bu durum ise test güvenliğini olumlu yönde etkiler (Thompson ve Weiss, 2011).

BOBUT uygulamalarında bireylerin ilk yetenek kestiriminden sonra hangi maddenin uygulanacağı da çeşitli yöntemlerle belirlenmektedir. Bu yöntemlerden maksimum Fisher bilgisi (Maximum Fisher Information-MFB), Owen'nın yaklaşık bayes yöntemi (Owen's Approximate Bayes Procedure) BOBUT uygulamalarının başlangıcından beri popüler olan iki yöntemdir (van der Linden, 1998; Weiss, 1982) aynı zamanda bu yöntemler klasik madde seçim yöntemleri olarak da sınıflandırılmaktadır. Kullback-Leibler bilgisi (Kullback-Leibler information-KLB) ve ağırlıklandırılmış olabilirlik bilgi ölçütü (likelihood-weighted information criterion-AOB) ise modern madde seçim yöntemleri olarak sınıflandırılmaktadır (van der Linden ve Pashley, 2000). Maksimum sonsal ağırlıklandırılmış bilgi (maximum posterior-weighted information-MSAB), maksimum beklenen bilgi (maximum expected information-MBB), minimum beklenen sonsal varyans (minimum expected posterior variance-MBSV) maksimum beklenen sonsal ağırlıklandırılmış bilgi (maximum expected posterior-weighted information-MBSAB) (van der Linden, 1998), a tabakalama ve verimlilik dengeli bilgi (VDB) kullanılan diğer madde seçim yöntemleridir (Han, 2012).

BOBUT uygulamalarında yetenek kestirimi için de kullanılan yöntemler bulunmaktadır. BOBUT uygulamalarının ilk yıllarından itibaren en sıklıkla kullanılan yetenek kestirim yöntemi maksimum olabilirlik (Maximum Likelihood – MO) yöntemidir (Hambleton vd, 1991; van der Linden ve Pashley 2000). Ağırlıklandırılmış olabilirlik kestirimi (weighted likelihood estimator, AOK), beklenen sonsal dağılım (expected a posteriori, BSD) ve maksimum sonsal dağılım (maximum a posteriori, MSD) sıklıkla kullanılan diğer yetenek kestirim yöntemleridir (van der Linden ve Pashley, 2000).

BOBUT uygulamalarının son aşamasında ise çeşitli sonlandırma kuralları bulunmaktadır. Bu uygulamalarda sabit madde sayılı kağıt-kalem testlerinden farklı olarak her birey farklı sayıda maddeye yanıt vermektedir. Bireyin yanıtladığı her maddeden sonra belirlenen test sonlandırma kuralına göre testin sonlanıp

sonlanmayacağına karar verilir (Hambleton vd., 1991). Bir BOBUT uygulaması, önceden belirlenmiş madde sayısına, önceden belirlenmiş standart hata seviyesine veya önceden belirlenmiş bir süreye ulaştığında sonlanır (Wainer, 2000). Bu sonlandırma kurallarından en yaygın olarak kullanılan standart hata yöntemidir. Bu yönteme göre yetenek kestirimi için önceden belirlenen bir standart hata değerine ulaştığında test sonlanmaktadır (Boyd vd., 2010).

Bir BOBUT uygulaması parametreleri kestirilmiş madde havuzu ve başlatma kuralını kabul ederek çalışır. Yetenek kestirim ve madde seçim yöntemi test sonlandırma kuralı karşılanana kadar sırayla uygulanır. Örneğin BOBUT uygulamasına katılacak bir bireye madde parametreleri önceden kestirilmiş madde havuzundan araştırmacı tarafından belirlenen bir başlatma yöntemiyle madde uygulanır. Madde cevaplandıktan sonra puanlanır ve bireyin yetenek kestirimi yapılır. Bireyin kestirilen yeteneğine uygun yeniden madde havuzundan madde seçilir. Bu süreç bir döngü halindedir. Sonlandırma kuralı karşılanana kadar birey maddelere yanıt verir (Thompson ve Weiss, 2011). Bu süreçte madde seçimi ve yetenek kestirimi anlık olarak gerçekleşir (van der Linden ve Pashley, 2010).

Geleneksel kağıt-kalem testlerinde testi alan bireyler kendi yetenek düzeyleri için çok zor veya çok kolay olan maddelere yanıt vermek için uzun zaman harcarlar ayrıca testteki tüm maddeleri cevaplamak zorundadırlar. Ancak BOBUT uygulamalarında bireylerin testteki tüm maddeleri yanıtlaması gerekmez (Hambleton ve Swaminathan, 1985). Tüm bireylerin aynı maddeleri yanıtladığı geleneksel kağıt-kalem testlerinin aksine BOBUT uygulamalarında birey yetenek düzeyine uygun zorlukta maddeleri yanıtlar. Bu durumda her bireyin yanıtladığı maddeler birbirinden farklı olur. Bireyin yetenek düzeyine uygun maddeler sorulduğu için daha az madde ile daha yüksek hassasiyette ölçümler yapılır (Segall, 2005). Ayrıca BOBUT uygulamalarının bilgi düzeyi yüksek maddelerin uygulanması nedeniyle daha az madde ile daha güvenilir ve geçerli sonuçlar vermesi ve bu sonuçlarını anında bildirmesi, daha iyi test güvenliği ve artan ölçüm hassasiyeti sağlaması, geleneksel yöntemlere göre avantaj sağlamaktadır (Wainer, 2000; Weissman, 2006).

Uluslararası yapılan çeşitli geniş ölçekli uygulamalarda ve ülkelerin kendi ulusal sınavlarında BOBUT uygulamaları yaygın olarak kullanılmaktadır. Fakat bu uygulamalar Türkiye'de yapılan ulusal çaptaki sınavlarda kullanılmamaktadır. Türkiye'de Öğrenci Seçme ve Yerleştirme Merkezi (ÖSYM) son birkaç yıldır Yükseköğretim Kurumları Yabancı Dil Sınavı (YÖKDİL), Yabancı Dil Bilgisi Seviye Tespit Sınavı (YDS) gibi

sınavları bilgisayar ortamına taşımış ve e-YÖKDİL, e-YDS adı altında uygulamaktadır. Bu uygulamalar BOBUT uygulaması şeklinde olmayıp sınavların kağıt-kalem formunun bilgisayar ortamına taşınmış halidir. Yani bu sınavların kağıt-kalem formu ile madde sayısı, süresi ve puan hesaplama yöntemi aynıdır. ÖSYM ve Milli Eğitim Bakanlığı (MEB) tarafından yapılan diğer birçok sınav, kağıt-kalem testi olarak yapılmakta ve bu sınavların sonucunda bireyler hakkında önemli kararlar verilmektedir. Bu sınavlarda bireylerden yetenek düzeyinin altında ve üstünde olan tüm maddelere yanıt vermeleri beklenmektedir. Aynı zamanda bu sınavlarda madde sayısı fazla olduğu için sınav için verilen süre de uzun olmaktadır. Ayrıca bu sınavlar için her yıl yüzlerce maddenin yazılması ve bu sınavların gizliliği gibi çeşitli sınırlılıklar bulunmaktadır. Bu sınırlılıklara rağmen ülkemizde kağıt-kalem testi uygulamalarının devam ettiği görülmektedir. MTK gibi güçlü varsayımları ve matematiksel modelleri olan bir kuram temelinde şekillenen ve birçok avantajı olan BOBUT uygulamalarının ülkemizde yapılan sınavlarda da kullanılmasının önemli olduğu düşünülmektedir.

Ülkemizde lisansüstü eğitime girişte, akademik personel atamalarında, yurtdışına lisansüstü eğitim için gönderilecek adayların seçiminde ve bunlara benzer amaçların gerçekleştirilmesinde kullanılmak üzere adayların sayısal ve mantıksal akıl yürütme (muhakeme) becerilerinin sayısal testi ile, sözel akıl yürütme (muhakeme) becerilerinin ise sözel testi ile ölçüldüğü Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) yapılmaktadır. ALES toplamda yüz, her bir alt testte elli madde içeren ve yılda üç defa yapılan bir sınavdır (ÖSYM, 2023). Bu sınavlarda her bir ALES için yeni maddeler yazılmaktadır çünkü testte yer alan maddelerin bir kısmı kamuoyu ile tamamı ise sınava giren adaylarla paylaşılmaktadır. Bu sınavların yılda üç kez yapıldığı düşünüldüğünde, sınavda ölçülen beceriler değişmediği halde aynı becerileri ölçen yeni madde yazmak hem zor hem de maliyetli olabilmektedir. Hâlbuki bu sınav BOBUT olarak uygulanırsa maddeler paylaşılmadığı ve parametreleri kestirilmiş geniş bir madde havuzu uygulama öncesinde hazırlandığı ve her birey yetenek düzeyine göre farklı maddeleri aldığından maddeler tekrar tekrar kullanılabilir. ALES'te birey yetenek düzeyine uygun olan olmayan tüm maddeleri yanıtlamak zorunda olduğu için birey çok fazla sayıda maddeyi yanıtlar. BOBUT uygulamalarında ise bireyin yeteneğine uygun maddeler kullanıldığı için daha az madde ile sınav sonlanır. Bir diğer nokta ise bu sınavlarda içerik ve kapsamın değişmemesine rağmen farklı sınav dönemlerinde uygulanan maddeler arasında güçlük düzeyi bakımından farklılıklar olduğu iddiasıdır. Oysaki bu sınav BOBUT olarak uygulansa birey yetenek düzeyine uygun maddeleri

yanıtlayacağı için farklı dönemlerde uygulanan maddelerin güçlük düzeylerindeki farklılıktan kaynaklı bireylerin yetenek kestirimine olan olumsuz etkisine yönelik sorun da çözülmüş olur. ALES, Amerika'da uzun bir süredir BOBUT olarak uygulanan GRE ve GMAT sınavları ile benzer bir amaca sahiptir. Bu sınavlar BOBUT uygulamalarının avantajlarından faydalanmaktadır ve ülkemizde yapılan ALES sınavının da bu avantajlardan yararlanması için BOBUT olarak uygulanmasının önemli olduğu düşünülmektedir.

BOBUT uygulamalarının kâğıt-kalem testlerine göre avantajları olsa da bazı dezavantajları da bulunmaktadır. BOBUT uygulamalarında maddeler tek tek bireyin önceki maddeye verdiği cevaba göre otomatik olarak gelir ve testin bütünü uygulama anında bireye özel otomatik olarak oluşur. Bireye özgü oluşan testlerin ilgili konu alanının kapsamını temsil etme durumunun uzmanlar tarafından gözden geçirilememesi, aynı zamanda bireyin maddeye tekrar dönememesi veya boş bırakamaması BOBUT uygulamalarının dezavantajlarındandır (Hendrickson, 2007; Lunz vd., 1992; Mead, 2006). Bu dezavantajları ortadan kaldırmak için son zamanlarda çok aşamalı bilgisayar ortamında bireye uyarlanmış testler (Multistage testing- ÇABOBUT) üzerine çalışmalar yoğunlaşmıştır (Çoban, 2020; Ertaş Polat, 2022; Karamese, 2022; Keyin, 2017; Yiğiter, 2023; Valdivia Medinaceli, 2023; Zheng vd., 2012; Wang, 2013; Wang, 2017). ÇABOBUT uygulamaları da BOBUT gibi bireyin yetenek düzeyine uygun olarak tasarlanmış bir test sistemidir. BOBUT uygulamalarında birey maddeleri tek tek yanıtlarken ÇABOBUT uygulamalarında ise yetenek düzeyine göre belirli sayıda ve özellikteki maddelerden oluşan madde setlerini yanıtlar. Modül diye adlandırılan bu setler sınav öncesinde önceden belirlenen hedef test bilgi fonksiyonlarına ve test belirtke tablosuna dayalı olarak güçlük düzeylerine göre ise kolay, orta ve zor modül olarak oluşturulmaktadır (Hendrickson, 2007; Luecht ve Sireci, 2011; Zheng vd., 2012). ÇABOBUT uygulamalarının da birtakım dezavantajı bulunmaktadır. BOBUT uygulamalarında bireye uygulanan her madde sonunda geçici yetenek düzeyi kestirilirken ÇABOBUT uygulamalarında ise bireye uygulanan her modül sonunda geçici yetenek düzeyi kestirilmektedir. Bu durum ise ÇABOBUT uygulamalarının daha düşük ölçme kesinliği sunduğunu göstermektedir (Patsula, 1999; Wang, 2017). Bir diğer dezavantajlı nokta ise bireyin yetenek düzeyine uygun olmayan bir modüle yönlendirilmesi bir başka deyişle yanlış yönlendirme (misrouting) yapılmasıdır. Yanlış yönlendirilen modülün bireyin gerçek yetenek düzeyinde daha az bilgi vermesi sonucu yeteneğin hatalı kestirilmesine neden olabilir. Bunun yanı sıra bireyin yanlış modüle yönlendirilmesinden

dolayı modülde yer alan maddeler bireyin yetenek düzeyine göre daha zor veya daha kolay maddeler olacaktır bu durum ise bireyi psikolojik açıdan etkileyecektir (Kim ve Moses, 2014). ÇABOBUT uygulamalarının BOBUT uygulamalarına göre testin kapsam dengelemesi açısından avantajlı olmasına karşın BOBUT uygulamalarında yeteneğin anlık kestirilmesi, yüksek ölçüm doğruluğu, daha kısa test uzunluğu ve yüksek test güvenliği açısından daha avantajlı olduğu belirtilmektedir (Yan vd., 2014).

Son on yılda, Uluslararası Öğrenci Değerlendirme Programı (PISA), Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) ve Uluslararası Yetişkin Becerilerinin Ölçülmesi Programı (PIAAC) gibi değerlendirmeler, kağıt-kalem testlerinden bilgisayar tabanlı testlere geçiş yapmıştır. Özellikle üst düzey yetenek ölçümleri için ölçüm hassasiyetini artırmak amacıyla, PIAAC ve PISA, sırasıyla 2016 ve 2018'de ÇABOBUT uygulamalarına geçmiştir (Valdivia Medinaceli, 2023). Bunun yanı sıra ETS tarafından yapılan GRE genel testi de (GRE general test) 2011 yılında ÇABOBUT'a geçmiştir (Hogan vd., 2016). GRE, ETS tarafından geliştirilen ve öğrencilerin lisansüstü programlara kabulü için bilgi ve hazırlık düzeylerini ölçen geniş ölçekli uluslararası bir sınavdır. GRE, GRE genel testi (GRE general test) ve GRE alan testi (GRE subject test) olmak üzere ikiye ayrılır (Educational Testing Service [ETS], 2018; ETS, 2019; Rankin, 2022). GRE genel testi sözel yetenek, sayısal yetenek ve analitik yazma bölümlerini içermektedir. GRE alan testi ise bireylerin Matematik, Fizik ve Psikoloji gibi alanlardaki akademik bilgi ve becerilerini değerlendirmek için tasarlanmış özgün bir sınavdır (ETS, 2025). GRE alan testi BOBUT olarak uygulanmaya devam ederken GRE genel testi ÇABOBUT uygulamalarına geçmiştir. Bu geçişin çeşitli nedenleri bulunmaktadır.

1952'den bu yana GRE genel testine katılan bireyler yetenek düzeyi, cinsiyet ve etnik köken/ırk açısından eskisine göre çok daha çeşitli hale gelmiştir. Testi alan bireylerin %30'undan fazlası Amerika Birleşik Devletleri (ABD) dışından gelmektedir (ETS, 2014). Popülasyonun değişmesi ile GRE genel testindeki sözel yetenek ve sayısal yetenek bölümlerinin ortalamalarının, ölçeğin ortasının (yani 500) dışına kaydığı tespit edilmiştir. Örneğin, ABD dışından gelen test alıcıların sayısı arttıkça, özellikle Asya'dan gelen katılımcılar, mühendislik ve bilim alanlarında daha fazla uzmanlaşma eğiliminde olduğu için sayısal bölüm de diğer katılımcılara göre daha iyi performans göstermiştir. Bu durumun sonucu olarak 760 ile 800 arasındaki yüksek puanları alan test alıcılarının sayısı artmıştır. Testi alan nüfusun değişmesi, madde türlerinin ölçülen becerilerle olan ilişkisi ve test puanlarının kullanımının genişlemesi gibi nedenlerden ötürü testin dikkatli bir şekilde incelenmesi gündeme gelmiştir. Böylelikle yeni içerik, yeni puanlama

ölçekleri, nicel karşılaştırmalar ve sayısal veri girişi gibi yeni soru türleri ile revize edilmiş GRE genel testi ÇABOBUT olarak uygulanmaya başlamıştır (ETS, 2014).

ALES'e benzer amaçla kullanılan GRE genel testi ÇABOBUT uygulamasına geçerken benzer bir amaçla kullanılan GMAT ise BOBUT olarak uygulanmaya devam etmektedir. GMAT 1953'ten beri ETS tarafından, işletme alanında lisansüstü eğitim almak isteyen adayların başvuruları için yapılan lisansüstü kabul sınavıdır. GMAT 2006 da ETS'ten ayrılarak Lisansüstü Yönetim Kabul Konseyi (Graduate Management Admission Council- GMAC) tarafından yapılmaya devam edilen standart bir testtir. GMAT, GRE ile benzer alt testlerden oluşmaktadır. Okuduğunu anlama, eleştirel düşünme ve cümle düzeltme gibi becerilerin ölçüldüğü sözel bölüm; problem çözme, geometri, veri yeterliği sayısal mantık ve akıl yürütme gibi becerilerinin ölçüldüğü sayısal bölüm, analitik yazma bölümü ve 2012'de eklenen akıl yürütme, tablo ve grafik yorumlama gibi becerileri ölçen akıl yürütme-muhakeme (Integrated Reasoning), bölümlerini içermekteydi (Bedsole, 2013; Rudner, 2012). GMAT'te içerik değişikliğine gidilmiş olup yeni içerik Temmuz 2024'te piyasaya sürülmüştür. Yeni GMAT versiyonunda, problem çözme becerilerini değerlendiren Nicel Muhakeme (Quantitative Reasoning) bölümü, eleştirel akıl yürütme ve okuduğunu anlama becerilerini değerlendiren Sözel Muhakeme (Verbal Reasoning) bölümü ve veri okuryazarlığı becerilerini değerlendiren Veri analizi ve yorumlama (Data Insights) bölümü yer almaktadır (GMAC, 2024). GMAT 1997'den beri BOBUT olarak uygulanmaktadır (GMAC, 2023; Rudner, 2012; Slater, 2001). Herhangi bir alt testten başlayabilen birey ilk önce orta güçlükte bir madde ile testte başlar. Bireyin maddeye verdiği cevap doğru ise bir sonraki madde daha zor, maddeye verdiği cevap yanlış ise daha kolay bir madde uygulanmaktadır. GMAT madde bazında adaptif test özelliği taşırken GRE modül bazında adaptif test özelliği gösterir ve modüllerin içindeki maddeler kağıt-kalem testlerinde olduğu gibi doğrusaldır (GMAC, 2023). GMAT'in madde havuzunda her yetenek düzeyindeki birey için kapsama uygun binlerce madde bulunmaktadır. GMAT Sayısal testi üç kategori altında sınıflandırmaktadır. Bunlar beceri alanı (veri yeterliliği veya problem çözme), içerik tabanlı (cebir, aritmetik beceriler veya geometri) ve uygulama (uygulamalı veya formül tabanlı) şeklindedir. GMAC, her bireyin belirtilen kategorilerde belirli bir sayıda madde alması gerektiğini belirtmektedir. Bunu yaparken van der Linden (2005) tarafından geliştirilen gölge test yaklaşımını (shadow-test approach -STA) kullanmakta bu sayede kapsam dengelemesi sağlanmaktadır. GMAT maddelerin aşırı kullanımını engellemek için madde kullanım sıklığı kontrolü

uygulamaktadır. Ayrıca GMAT'te düzenli olarak madde yanlılığı ve madde parametre kayması da incelenmektedir (Rudner, 2009).

Türkiye'de GRE ve GMAT ile benzer bir amaçla uygulanan ALES'in, sayısal ve sözel becerileri ölçen alt testlerden oluşması açısından bu sınavlara benzerdir. Bu sınavlardan GMAT BOBUT uygulamalarının avantajlarını yararlanırken GRE ÇABOBUT uygulamalarının avantajlarından yararlanarak uygulanmaktadır. GRE genel sınavının analitik yazma bölümü içermesi, farklı madde türlerini barındırması, farklı puanlama ölçeklerini bulundurması ve farklı kültürlerden adayların katıldığı uluslararası geniş ölçekli bir sınav olması açısından ALES'ten farklılaşmaktadır. ALES'in sayısal alt testinin adayların sayısal ve mantıksal akıl yürütme (muhakeme); sözel alt testinin sözel akıl yürütme (muhakeme) becerilerini ölçmeye yönelik (ÖSYM, 2018a) ulusal çapta bir sınav olması, tek tür maddelerden oluşması, sınavı alan popülasyonun ciddi bir oranının GRE genel testi gibi farklı milletlerden oluşmaması ayrıca BOBUT uygulamalarının ÇABOBUT'tan daha hassas ölçümler yaptığı (Çoban, 2020; Hambleton ve Xing, 2006; Keng, 2008; Keyin, 2017; Kim, 2010; Kim vd., 2012; Wang, 2017) sonucuna ulaşılmasından dolayı bu araştırmada GMAT'inde yararlandığı BOBUT uygulaması kullanılmasına karar verilmiştir.

Alan yazın incelendiğinde bilgisayar ortamında bireye uyarlanmış test ile ilgili birçok çalışma bulunmaktadır. Bunlar çok kategorili maddelerin yer aldığı çeşitli envanter ve ölçekler ile yürütüldüğü (Alkan, 2021; Aybek, 2016; Gibbons vd., 2016; Demir, 2018; Jensen, 2017; Oswald vd., 2015; Petersen vd., 2016; Scullard, 2007; Smits vd., 2011; Şimşek, 2017) çalışmalar; ikili puanlanan maddelerin yer aldığı (Bulut ve Kan, 2012; Costa vd., 2009; Çoban, 2020; Han, 2009; Liu vd., 2013; Sayman Ayhan, 2015; Ogunjimi vd., 2021; Yen vd., 2012) çalışmalar ve bunun yanı sıra hem ikili hem de çok kategorili maddelerin yer aldığı (Keng, 2008; Kim, 2010) çalışmalar bulunmaktadır.

İkili puanlanan maddelerin yer aldığı yüksek riskli sınavların farklı koşullar altında BOBUT olarak incelendiği çeşitli araştırmalar bulunmaktadır. Çoban (2020) yaptığı araştırmada, YÖKDİL'in bilgisayar ortamında bireye uyarlanmış test, çok aşamalı bireyselleştirilmiş test (ÇABOBUT) ve bilgisayarlı sınıflama testi (BST) bireyselleştirilmiş test yaklaşımlarından hangisinin daha uygun olduğunu simülasyon ortamında karşılaştırmıştır. BOBUT uygulamasında yetenek kestirim yöntemi olarak beklenen sonsal dağılım (BSD), madde seçim yöntemi olarak maksimum Fisher bilgisi (MFB) ve sonlandırma kuralı olarak sabit uzunluk yöntemi kullanılmıştır. Araştırmada BOBUT'un ÇABOBUT ve BST'ye göre ölçme kesinliği açısından daha iyi olduğu

sonucuna ulaşmıştır. Ogunjimi vd. (2021), Nijerya da yüksek riskli sınavların değerlendirilmesinde kullanılmak üzere bir BOBUT uygulaması araştırması yürütmüştür. Madde parametre kestirimi için 3PL model kullanılan uygulamada yetenek kestirim yöntemi olarak maksimum olabilirlik (MO), madde seçim yöntemi olarak MFB, sonlandırma kuralı olarak hem sabit hem de değişen test sonlandırma kuralı kullanılmıştır. Araştırmada sabit uzunluklu test sonlandırma kuralının yetenek kestiriminde daha hassas ölçümler yaptığı belirtilmiştir. Bulut ve Kan (2012) 2008 yılı ALES sayısal 1, sayısal 2 ve sözel alt testlerini BOBUT olarak uygulanabilirliğini gerçek veriye dayalı post hoc simülasyonlarıyla incelemişlerdir. Simülasyon koşullarında yetenek kestirim yöntemi olarak MO, BSD, ağırlıklı en küçük kareler ve maksimum sonsal dağılım (MSD) madde seçim yöntemi olarak MFB, sonlandırma kuralı olarak da değişen uzunluk yöntemi kullanılmıştır. Araştırmada yetenek kestirimi için BSD'nin daha iyi olduğu, en doğru yetenek kestiriminin sayısal 1 alt testinden elde edildiği ve BOBUT'un testte yer alan madde sayısının %44-%88 oranında azalttığı sonucuna ulaşılmıştır. Imai vd., (2009) tarafından yapılan araştırmada Japonca'yı yabancı dil olarak ölçmek için geliştirilen Japon Bilgisayar Destekli Test (J-CAT) özellikleri incelenmiştir. Testin orta güçlükte bir maddeyle başladığı, yetenek kestirimde BSD, madde seçim yöntemi olarak MFB kullanıldığı ve test sonlandırma kuralı olarak sabit uzunluklu veya değişen uzunluklu olabildiğini, testin adayın belirlenen durdurma kurallarından herhangi birine ulaştığında durduğunu belirtmiştir. Molina (2009) İngilizce öğrenimi için geliştirdiği bilgisayar destekli kelime dağarcığı testi ile İngilizce yeterliliğini kağıt-kalem testi ve BOBUT uygulamasıyla karşılaştırmıştır. Madde parametre kestirimi için 3PL model kullanılan uygulamada yetenek kestirim yöntemi olarak MO, madde seçim yöntemi olarak MFB, sonlandırma kuralı olarak hem sabit hem de değişen test sonlandırma kuralı kullanılmıştır. Araştırmada BOBUT'un test uzunluğunu kısaltmada ve ölçme hassasiyetini arttırmada kağıt-kalem testi yerine kullanabilirliğini önermiştir. Sayman Ayhan (2015) ise yükseköğretime giriş sınavının fen alt testinin BOBUT olarak uygulanabilirliğini incelediği araştırmasında yetenek kestirim yöntemi olarak MO ve BSD, sonlandırma kuralı olarak da sabit ve değişen uzunluk yöntemi kullanmıştır. Araştırmanın bulgularına göre, yetenek kestirim yöntemi olarak BSD'nin tüm koşullarda MO'ya göre daha iyi olduğu, testin sonlanması için BSD'nin MO'ya göre daha fazla madde gerektirdiği sonucuna ulaşılmıştır. Bu çalışmada ise alan yazındaki yüksek riskli sınavla çalışan araştırmalardan farklı olarak yetenek kestirim yöntemi olarak MO ve BSD; madde seçim yöntemi olarak MFB, Kullback-Leibler bilgisi (KLB), verimlilik

dengeli bilgi (VDB), ağırlıklandırılmış olabilirlik bilgi kriteri (AOB); sonlandırma kuralı olarak hem sabit hem de değişen uzunluk yöntemleri 40 farklı koşul altında incelenecektir. Çalışmalarda test sonlandırma kuralı olarak yaygın olarak değişen uzunluklu yöntemler kullanıldığı (Alkan, 2021; Aybek, 2016; Bulut ve Kan, 2012; Çoban, 2020) bunun yanı sıra hem değişen hem de sabit uzunluklu test sonlandırma kuralının kullanıldığı (Babcock ve Weiss 2012; Demir, 2018; Imai vd., 2009; Kaya, 2022; Sayman Ayhan, 2015; Sulak, 2013) çalışmalarda bulunmaktadır. BOBUT uygulamalarında her bireyin farklı sayıda ve farklı maddeleri alıyor oluşu özellikle üniversiteye giriş gibi yüksek riskli sınavlarda toplum tarafından tepki alabileceğinden, BOBUT uygulamalarının daha az riskli bir sınav olan ALES'te denenmeye başlamasının bu açıdan da önemli olduğu düşünülmektedir. Bu nedenle, bu çalışmada ALES'te ikili puanlanan maddelerde sabit ve değişen uzunluklu test sonlandırma kuralı ile farklı madde seçim yöntemlerinin karşılaştırılıp en uygun BOBUT koşullarının neler olduğunun incelenmesine gereklilik görülmüştür.

Türkiye'de yapılan yüksek riskli sınavlarla ilgili BOBUT çalışmalarında, BOBUT'un her aşaması için genellikle tek bir yöntem kullanıldığı görülmüş ve BOBUT'un her aşaması için farklı yöntemlerin karşılaştırıldığı çalışmaya rastlanılmamıştır. Ayrıca benzer çalışmalarda canlı BOBUT uygulaması yapılan sınırlı (Kalender, 2011) çalışmaya rastlanılmıştır. Kalender (2011) yaptığı canlı BOBUT uygulamasında araştırma kapsamında Öğrenci Seçme Sınavı (ÖSS) 2007 yılı fen alt boyutunu incelemiştir. Bu çalışmada ise farklı olarak ALES'in tüm alt testleri için post-hoc simülasyonları ve canlı BOBUT uygulaması yürütülmüştür. Bu nedenlerden dolayı bu araştırmada ALES'te BOBUT için farklı yöntemler altında 40 koşul karşılaştırılması ve en uygun BOBUT koşullarının belirlenmesi, aynı zamanda canlı bir BOBUT uygulamasının yapılarak ALES için en uygun BOBUT koşullarının belirlenmesine ihtiyaç duyulmuştur.

Amaç

Bu araştırmanın amacı, kağıt-kalem testi olarak uygulanan ALES'in BOBUT olarak uygulanabilirliğini farklı simülasyon koşulları altında (yetenek kestirimi, test sonlandırma kuralı, madde seçim yöntemi) ve gerçek veri uygulamasına dayalı olarak araştırmaktır. Ayrıca simülasyon sonucu elde edilen en uygun BOBUT koşullarının canlı bir ALES BOBUT uygulaması ve ALES kağıt-kalem formu tarafından kestirilen yetenek

düzeylerinin doğruluğunu incelemektir. Bu amaçla aşağıda verilen alt amaçlar incelenecektir.

1. ALES madde havuzunda yer alan maddelerin MTK'ya göre parametreleri nasıldır?

2. Madde kullanım sıklığı ve kapsam dengeleme yöntemleri tüm koşullarda sabit tutulduğunda,

- yetenek kestirim yöntemi olarak iki farklı yöntem (Beklenen Sonsal Dağılım [BSD], Maksimum Olabilirlik [MO]),
- test sonlandırma kuralı olarak beş farklı yöntem (Standart hata SH<0.30, < 0.40 ve <0.50, sabit madde sayısı MS=15, MS=20) ve
- madde seçim yöntemi olarak dört faklı yöntem (Maksimum Fisher bilgisi, Kullback-Leibler Bilgisi, Ağırlıklandırılmış Olabilirlik Bilgi, Verimlilik Dengeli Bilgi) kullanıldığında;

ALES post-hoc simülasyonlarında uygulanan ortalama madde sayısı ve madde havuzu kullanım oranları, ortalama standart hata değerleri, RMSE, yanlılık ve Marjinal güvenirlik değerleri, ALES post-hoc simülasyonu ve ALES asıl formdan elde edilen yetenek düzeyleri (θ) arasındaki ilişki nasıldır?

3. ALES post-hoc simülasyonları sonucunda alt testler için ölçme hassasiyeti açısından en uygun BOBUT koşulları nedir?

4.ALES post-hoc simülasyon sonuçlarına göre belirlenen en uygun yetenek kestirim yöntemi, test sonlandırma kuralı ve madde seçim yöntemi kullanılarak oluşturulan ALES canlı BOBUT olarak uygulandığında:

a) Katılımcıların ALES kağıt-kalem formundan ve ALES canlı BOBUT uygulamasından kestirilen yetenek parametreleri arasındaki ilişki ne düzeydedir?
b) Katılımcıların ALES kağıt-kalem formundan ve ALES canlı BOBUT uygulamasından kestirilen yetenek parametreleri arasında manidar farklılık var mıdır?
c) Katılımcılara ALES canlı BOBUT uygulamasında uygulanan maddelerin içerik dağılımları ne düzeydedir?

Önem

Türkiye'de her yıl ÖSYM ve MEB tarafından seçme ve yerleştirme amaçlı birçok sınav yapılmaktadır. Bu sınavlar birey için önemli kararların verildiği yüksek riskli sınavlar olarak adlandırılmaktadır. Birey hakkında karar verilen bu önemli sınavlarda bireyin gerçek yeteneğinin veya puanının kestirilmesi önem arz etmektedir. Bu nedenle, bu sınavlardan elde edilen puanların daha doğru ve daha az hatalı olmasına yönelik yapılan çalışmaların alan yazına ve ilgili kurumlara önemli katkıları olacağı düşünülmektedir.

ÖSYM tarafından yapılan sınavlara katılan çok fazla öğrenci olduğu düşünüldüğünde hem sınavların maliyeti hem de sınavların önemi daha ön plana çıkmaktadır. Sınavların BOBUT olarak uygulanması, gizli bir madde havuzunun olması maddelerin tekrar tekrar kullanılması maliyeti azaltmaya yardımcı olabilir. Ayrıca bu sınavların taşınması, uygulanması, gözetmenliği, ifşası gibi birtakım lojistik ve güvenlik sorunları da bulunmaktadır. Bu sınavların BOBUT olarak uygulanması bu sorunları minimum seviyeye inmesine yardımcı olacaktır. Bu çalışmanın ALES testinin BOBUT olarak uygulanabilirliğini test ederek sınavın ekonomiklik, lojistik ve güvenlik sorunlarına çözüm sunması açısından önemli olduğu düşünülmektedir.

Türkiye'de bilgisayar ortamında bireye uyarlanmış test ile ilgili çalışmalar incelendiğinde yüksek riskli sınavlarla ilgili yapılan araştırmalar (Bulut ve Kan, 2012; Çoban, 2020; Kalender, 2011) sınırlı sayıdadır. Bu nedenle bireyler için önemli kararlar alınan yüksek riskli bir sınav olan ALES sınavının bilgisayar ortamında bireye uyarlanmış test olarak uygulanabilirliğinin incelenmesi önemli bulunmuştur. Bu amaçla bu araştırmanın bilişsel özelikleri ölçen BOBUT uygulamaları için alan yazına katkı sunacağı düşünülmektedir.

Dünya'da BOBUT olarak uygulanan GRE, GMAT, TOEFL, J-CAT gibi sınavlarda bireyin yetenek düzeyine uygun daha az madde ile istenilen ölçme kesinliğinde ölçmeler yapıldığı, bireyin yetenek düzeyine uygun sorular sorulduğu için motivasyon düşüklüğü yaşanmadığı ayrıca sınav anında otomatik veri toplanması ve sınav sonunda bireyin yetenek düzeyinin belirlenmesi gibi anında geri bildirim verildiği düşünüldüğünde (Hambleton vd., 1991; Ho, 2010; Linacre, 2000) ülkemizde de bu uygulamalara geçilmesinin sağlayacağı avantajların önemli olduğu düşünülmektedir. Ülkemizde uygulanan yüksek riskli sınavların BOBUT uygulamaları olarak uygulanması durumunda bu araştırmanın bulgularının faydalı olacağı umulmaktadır.

Sınırlılık

Bu araştırma;

Madde seçim yöntemi olarak seçilen Maksimum Fisher bilgisi, Kullback-Leibler Bilgisi, Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, Verimlilik Dengeli Bilgi,

Yetenek kestirim yöntemi olarak seçilen maksimum olabilirlik ve beklenen sonsal dağılım,

Sonlandırma kuralı olarak beş farklı yöntem (SH<0.30, SH< 0.40 ve SH<0.50, MS=15, MS=20),

ALES 2021 yılı üç döneme ait 10.000 uygulama verisi ile sınırlıdır.

Tanımlar ve Kısaltmalar

KK: Kâğıt-Kalem Testi

MTK: Madde Tepki Kuramı

KTK: Klasik Test Kuramı

BOBUT: Bilgisayar Ortamında Bireye Uyarlanmış Test

MFB: Maksimum Fisher Bilgisi

KLB: Kullback-Leibler Bilgisi

AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri

VDB: Verimlilik Dengeli Bilgi

MO: Maksimum Olabilirlik

BSD: Beklenen Sonsal Dağılım

ALES: Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı

ÖSYM: Öğrenci Seçme ve Yerleştirme Merkezi

MEB: Milli Eğitim Bakanlığı

GRE: Lisansüstü Eğitim Giriş Sınavı

GMAT: İşletme ve Yönetim Programları Lisansüstü Eğitim Giriş Sınavı

1PL: Bir Parametreli Lojistik Model

2PL: İki Parametreli Lojistik Model

3PL: Üç Parametreli Lojistik Model

BÖLÜM 2

KURAMSAL ÇERÇEVE VE İLGİLİ ARAŞTIRMALAR

Bu bölümde, Madde Tepki Kuramı ve BOBUT kavramlarına ilişkin kuramsal bilgiler yer almaktadır. BOBUT uygulamaları MTK temelli olarak uygulandığı için ilk olarak MTK ve varsayımlarından bahsedilmiş devamında iki kategorili ve çok kategorili MTK modelleri açıklanmıştır. Daha sonra BOBUT uygulamalarının özellikleri anlatılmıştır. Bu kapsamda BOBUT'un aşamaları olan madde seçim yöntemleri, yetenek kestirim yöntemleri ile test sonlandırma kuralları anlatılmıştır. Son olarak yurtiçinde ve yurtdışında yapılan ilgili araştırmalara yer verilmiştir.

Madde Tepki Kuramı

Test teorisi klasik test kuramı ve madde tepki kuramı olmak üzere iki ana kuram çerçevesinde şekillenmektedir. Klasik Test Kuramı (KTK) 20. yüzyılın büyük bir bölümünde psikolojik test geliştirmenin temelini oluşturmuştur. KTK'nın temelleri Spearman'ın (1907, 1913) çalışmalarına kadar dayanmaktadır. Lord ve Novick'in (1968) Madde Tepki Kuramı (MTK) üzerindeki çalışmalarını takiben psikolojik ölçümler için temel bir yaklaşım olarak hızla popüler hale gelmiştir (Embretson ve Reise, 2000). KTK'da bireyin yetenek düzeyi bir testteki gerçek puan kavramına dayanarak belirlenir. Madde Tepki Kuramı'nda ise bireyin test performansının yeteneği tarafından tahmin edilebileceği fikrine dayanır. KTK'nın temeli bireyin gerçek ve gözlenen puanına dayanır ama MTK'nın temeli yetenek puanlarıdır. İkisini ayıran özellik test bağımlılığıdır. KTK'da gerçek puanlar teste bağlı iken MTK'da gerçek puan testten bağımsızdır. Örneğin zor bir testte bireyin gerçek puanı düşük, kolay bir testte ise yüksek olabilir. Bununla birlikte MTK ile ilişkilendirilen bireyin yetenek puanı testin zorluk seviyesinden bağımsız olarak sabittir. Test zor da olsa kolay da olsa bireyin yetenek düzeyi aynıdır değişmez. MTK bireyin madde performansı ile yeteneği arasındaki ilişkiyi Madde Karakteristik Eğrisi (MKE) olarak adlandırılan monoton artan bir fonksiyon yardımıyla açıklar. MKE'ye göre bireyin yetenek seviyesi arttıkça, maddeye doğru cevap verme olasılığı da artar (Hambleton ve Swaminathan, 1985; Hambleton vd., 1991). Düşük yetenek düzeyindeki bireylerin maddeyi doğru yanıtlama olasılığı düşük olup 0'a

yakınken yüksek yetenek düzeyindeki bireylerin ise maddeyi doğru yanıtlama olasılığı artar ve 1'e yaklaşır. Her bir maddenin ayrı bir MKE'si bulunur ve MTK'da tanımlı modeller bu eğri temelinde şekillendiği için oldukça önemli bir fonksiyondur (Baker, 2001).

Gerçek puan teorisine dayalı olarak KTK birtakım varsayımlara dayanmaktadır. Bu varsayımlara göre gerçek puan ve hata arasında ilişki yoktur, farklı test formları arasındaki hatalar birbirinden ilişkisizdir ve bir test formundaki hata başka bir test formundaki gerçek puan ile ilişkisizdir. Bu varsayımlar zayıf varsayım olarak kabul edilir. Buna karşın MTK'da güçlü varsayımlar bulunmaktadır (de Ayala, 2009). Bu varsayımlar tek boyutluluk (unidimensionality) ve yerel bağımsızlık (local independence) gibi iki temel varsayım olarak kabul edilmektedir (Hambleton vd., 1991). Tek boyutluluk varsayımı, ölçme aracındaki maddelerin tamamının tek bir özelliği veya boyutu ölçülmesi olarak tanımlanmaktadır (Han ve Hambleton, 2014; Hambleton ve Swaminathan, 1985). Bu varsayımı sağlamak her zaman mümkün olmayabilir, bunun için baskın bir boyutun veya faktörün olması tek boyutluluk varsayımını sağlar. Tek boyutluluk varsayımının her zaman kontrol edilmesi gerekir ve her zaman aynı özelliğin ölçüldüğünden emin olunması gerekir (Hambleton ve Swaminathan, 1985). Tek boyutluluğu sağlamak için ölçülmek istenen yapının çok iyi tanımlanması akabinde yapıya uygun madde yazılması gerekir. Tek boyutluluk varsayımını test etmek için faktör analizi özdeğer ve yamaç birikinti grafiğinin incelenmesi, Shout tek boyutluluk testi ve artık analizi yapılabilmektedir (deMars, 2010). MTK'nın bir diğer önemli varsayımı ise yerel bağımsızlıktır. Bu varsayıma göre yetenek parametresi kontrol altına alındığında maddelere verilen yanıtlar istatistiksel olarak birbirinden bağımsız olduğu kabul edilir. Başka bir ifadeyle bireyin bir testte yer alan maddeye verdiği cevabın diğer maddelere nasıl cevap verdiğinden bağımsız olmasıdır (de Ayala, 2009; Embretson ve Reise, 2000; Hambleton vd., 1991). Maddelerin yerel bağımsızlığını kontrol etmek için Yen (1984) tarafından geliştirilen Q3 istatistiği sıklıkla kullanılan bir yöntemdir (deMars, 2010).

MTK modelleri veri yapısına göre iki kategorili (dichotomous) ve çok kategorili (polytomous); boyut yapısına göre ise tek boyutlu (unidimensional) ve çok boyutlu (multidimensional) olarak sınıflandırılmaktadır (Mair, 2018). Bu çalışmada tek boyutlu MTK modellerine ait iki ve çok kategorili modeller hakkında bilgi verilmiştir.

İki Kategorili Madde Tepki Kuramı Modelleri

MTK tarihinde ele alınan ilk madde yapısı iki kategorili maddeler olmuştur (van der Linden ve Hambleton, 1997). İki kategoriye sahip maddeler genellikle 1 ve 0 olarak kodlanan doğru ve yanlış olarak puanlanan test maddelerdir. İki kategorili maddeler için yaygın kullanılan modeller 1 Parametreli Lojistik Model (1 PL), 2 Parametreli Lojistik Model (2 PL) ve 3 Parametreli Lojistik Model (3 PL)'dir. Bu modeller her bir madde için doğru cevap verme olasılığı ile bireyin yeteneği arasındaki ilişkiyi gösteren fonksiyonda kullanılan madde parametre sayısına göre adlandırılır (Embretson ve Reise, 2000; Hambleton vd., 1991). Doğru bir yanıtın, yanlış bir yanıttan daha yüksek bir yetenek düzeyini temsil ettiği varsayılmıştır, bu nedenle bireyler ve maddeler arasındaki etkileşimleri modellemek için monotonik olarak artan matematiksel fonksiyonlar kullanılmıştır (Reckase, 2009). Bu iki kategorili MTK modelleri, testte yer alan maddelerin güçlük, ayırt edicilik ve tahmin etme (şans) gibi özellikleri dikkate alınarak parametre sayılarına göre oluşturulmuştur (van der Linden ve Hambleton, 1997). Bu modeller arasındaki temel fark maddeleri tanımlarken kullanılan parametre sayısıdır. Yaygın kullanılan iki kategorili MTK modelleri bir, iki ve üç parametreli lojistik modellerdir ve bu modeller içerdikleri parametre sayısından dolayı böyle adlandırılmaktadır.

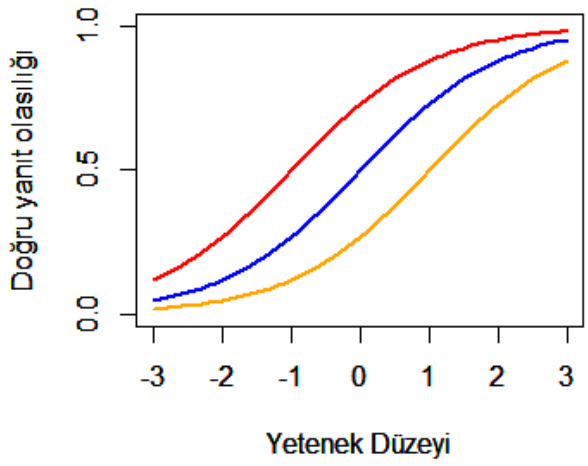
1 Parametreli Lojistik Model- 1 PL. 1 parametreli lojistik modelde bireyin yetenek düzeyi hesaplanırken sadece maddelerin güçlük parametresi (b) temel alınmaktadır. "b" parametresi bir maddeye doğru cevap verme olasılığının %50 olduğu noktadaki yetenek düzeyidir. Bu parametre genelde -3 ile +3 arasında yer alır ve bu parametre büyüdükçe soru zorlaşır, küçüldükçe ise soru kolaylaşır. Farklı maddeler farklı güçlük parametrelerine sahiptir (Baker, 2001; Embretson ve Reise, 2000; Hambleton vd., 1991). Bu modelde ayırt edicilik parametresi (α) tüm maddeler için eşit kabul edildiği için bireyin yeteneğinin hesaplanma sürecinde etkisi yoktur. 1 PLM aynı zamanda Rasch model olarak da bilinmektedir. 1 PL'de α parametresi herhangi bir değere sabitlenirken, Rasch modelde α parametresi "1" olacak şekilde sabitlenmektedir (Embretson ve Reise, 2000). Şans parametresi (c) ise tüm maddeler için sıfıra sabitlenmektedir. Bu modele ilişkin matematiksel eşitlik (Hambleton vd., 1991) ve madde karakteristik eğrisi aşağıda verilmiştir.

$$P_i(\theta) = \frac{e^{(\theta - bi)}}{1 + e^{(\theta - bi)}} \qquad i = 1, 2, \ldots, n \text{ (Eşitlik 1)}$$

$P_i(\theta)$: θ yetenek düzeyinde olan bir bireyin i maddesini doğru cevaplaması olasılığıdır.

bi =i maddesinin güçlük parametresidir.

n= testteki madde sayısı



Şekil 1. 1 Parametreli Lojistik Modele Ait Madde Karakteristik Eğrisi

Şekil 1 incelendiğinde sarı ile belirtilen maddenin doğru yanıtlama olasılığının %50 olduğu noktadaki yetenek düzeyi "1" olarak ölçülmüştür. Bu maddenin güçlüğü b=1 olup yetenek düzeyi "1" olan bireyler için uygun bir maddedir. Benzer şekilde mavi ile belirtilen maddenin b parametresi "0" kırmızı ile belirtilen madde ise "-1" olarak ölçülmüştür. Ayırt edicilik parametresi ise tüm maddeler için aynı olduğu için eğriler çakışmamaktadır.

2 Parametreli Lojistik Model- 2 PL. 2 parametreli lojistik modelde, bireyin yetenek kestirimi hesaplanırken maddenin güçlük parametresine ek olarak madde ayırt edicilik parametresi de modele dahil edilir. Yani modelde hem b hem de α parametresi birlikte bulunmaktadır (Embretson ve Reise, 2000). "α" parametresi maddenin MKE'de θ noktasındaki eğimini tanımlar ve teorik olarak (-∞, +∞) aralığında değer alır. Negatif ayırt ediciliğe sahip maddeler testten çıkarılır çünkü bireyin yetenek düzeyi arttıkça maddeye doğru cevap verme olasılığı azalıyorsa madde de bir hata vardır. Genellikle [0,

2] arasında değer alır ve 2'ye yaklaştıkça eğim dikleşir ayırt edicilik artar, 0'a yaklaştıkça eğim yataya yakınlaşır ve ayırt edicilik azalır. Bu modele ilişkin matematiksel eşitlik aşağıda verilmiştir (Hambleton vd., 1991).

$$P_i\ (\theta) = \frac{e^{D\alpha i(\theta - bi)}}{1+e^{D\alpha i(\theta - bi)}} \qquad i= 1, 2, \ldots, n \text{ (Eşitlik 2)}$$

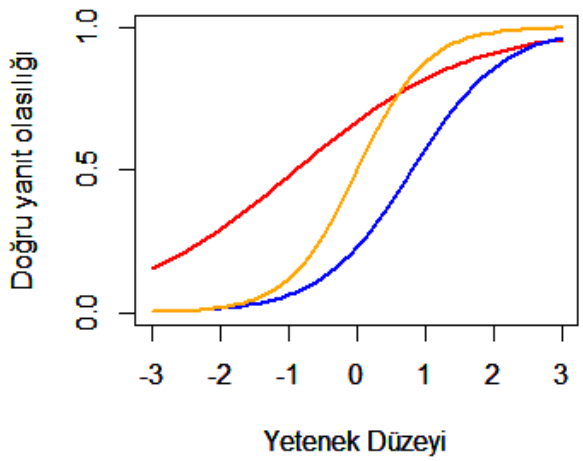
$P_i\ (\theta)$: θ yetenek düzeyinde olan bir bireyin *i* maddesini doğru cevaplaması olasılığıdır

bi =*i* maddesinin güçlük parametresidir

αi = *i* maddesinin ayırt edicilik parametresidir

n= testteki madde sayısı

D= 1.702 ölçekleme katsayısı



Şekil 2. 2 Parametreli Lojistik Modele Ait Madde Karakteristik Eğrisi

Şekil 2'de görüldüğü gibi 1PL modelden farklı olarak 2 PL modelde a parametreleri farklılaştığı için madde karakteristik eğrilerinin çakışması söz konusudur.

3 Parametreli Lojistik Model- 3 PL. 3 parametreli lojistik modelde, bireyin yetenek kestirimi hesaplanırken maddenin güçlük ve madde ayırt edicilik parametresine ek olarak tahmin etme (şans) parametresi de modele dahil edilir. Yani modelde b, α ve c parametresi birlikte bulunmaktadır. "c" parametresi düşük yetenekteki bireyin maddeyi

doğru cevaplama olasılığını temsil eder ve MKE için sıfıra eşit olmayan bir alt asimptot sağlar. Bu modele ilişkin matematiksel eşitlik aşağıda verilmiştir (Hambleton vd., 1991).

$$P_i(\theta) = ci + (1 - ci)\frac{e^{D\alpha i(\theta - bi)}}{1 + e^{D\alpha i(\theta - bi)}} \qquad i = 1, 2, \ldots, n \text{ (Eşitlik 3)}$$

$P_i(\theta)$: θ yetenek düzeyinde olan bir bireyin i maddesini doğru cevaplaması olasılığıdır

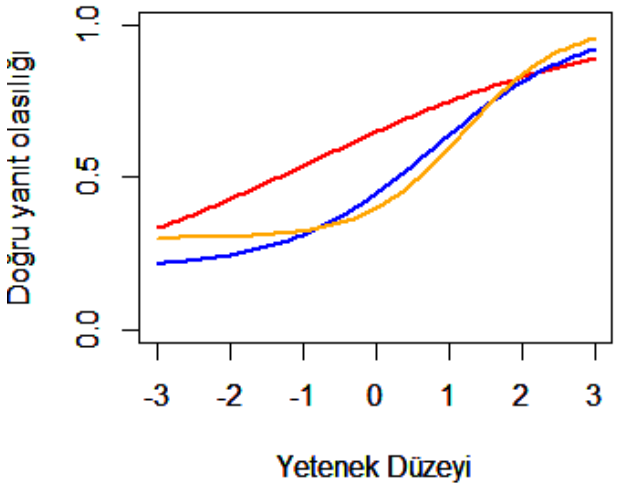
bi =i maddesinin güçlük parametresidir

αi = i maddesinin ayırt edicilik parametresidir

ci = i maddesinin tahmin etme parametresidir

n= testteki madde sayısı

D= 1.702 ölçekleme katsayısı



Şekil 3. 3 Parametreli Lojistik Modele Ait Madde Karakteristik Eğrisi

Şekil 3'te görüldüğü gibi 3 PL modelde 3 maddeye ilişkin madde karakteristik eğrilerinin başlangıç noktaları 1PL ve 2PL modelden farklı olarak değişmektedir.

MTK'nin avantajlarının sağlanabilmesi için öncelikli olarak test verilerine en uygun MTK modelinin seçilmesi gerekir (Hambleton ve Swaminathan, 1985). Hangi modelin veriye en uygun model olduğunu belirlerken ölçme aracının özellikleri, model veri uyumu, madde parametre kestiriminin amacı gibi birçok istatistik göz önünde bulundurulur (Embretson ve Reise, 2000). Örneğin, çoktan seçmeli maddelerde şans

faktörünü de dikkate alan 3 PL en yaygın kullanılan modeldir. Fakat güçlü çeldiricilerin olduğu maddelerde ise daha basit bir model olan 2 PL'nin de kullanılabileceği belirtilmektedir (deMars, 2010).

Çok Kategorili Madde Tepki Kuramı Modelleri

MTK modelleri ile yapılan çoğu çalışma, 0 ve 1 gibi ikili puanlanan test maddelerinden gelen verilerin analizine odaklanmış olsa da ikiden fazla kategoriye sahip maddeler için de çeşitli modeller geliştirilmiştir. Bu modeller, bireylerin açık uçlu maddelere, dereceli yanıtlar içeren anket ve ölçek maddelerine verdikleri yanıtları kestirmek için tasarlanmıştır. Bu tür maddelerde ikili puanlanan maddeler gibi sadece doğru veya yanlış cevap durumu olmadığı için birey-madde etkileşimini açıklayan MTK modelleri de farklıdır (Reckase, 2009). Maddeler iki yanıt kategorisinden fazla kategori içerir ve bu kategoriler kendi içinde sıralanmıştır. Örneğin likert tipi bir ölçekte, bir maddeye "kesinlikle katılıyorum", "katılıyorum", "katılmıyorum", "kesinlikle katılmıyorum" şeklinde olası dört yanıt kategorisi kullanarak yanıt verilebilir. Bu yanıt kategorileri ve dolayısıyla bireyin yanıtı, olumlu bir tavrı olumsuza doğru temsil eden bir ölçekte sıralanmıştır (de Ayala, 2009). Çok kategorili puanlanan maddeler için geliştirilen modellerin çoğu iki kategorili maddeler için geliştirilen modellerin genelleştirilmiş halidir (Han ve Hambleton, 2014). Tablo 1'de modellerin de Ayala (2009) tarafından yapılan sınıflandırılması ve hangi araştırmacılar tarafından geliştirildiği parantez içerisinde belirtilmiştir.

Tablo 1

Çok Kategorili Madde Tepki Kuramı Modelleri

Sıralı Puanlanan Modeller		Kategorik Puanlanan Modeller
Rasch Temelli Modeller	Rasch Temelli Olmayan Modeller	
Kısmi Puan Modeli (KPM [Partial Credit Model (Masters,1982)])	Genelleştirilmiş Kısmi Puan Modeli (GKPM [Generalized Partial Credit (Muraki,1992)])	Sınıflamalı Tepki Modeli (STM [Nominal Response Model (Bock,1972)])
Dereceleme Ölçeği Modeli (DÖM [Rating Scale Modeli (Andrich,1978; Anderson, 1977)	Aşamalı Tepki Modeli (ATM [Graded Response Model (Samejima,1969)])	Çoklu Seçenek Modeli (ÇSM [Multiple Choice Model (Bock–Samejima, 1979)])

Sıralı puanlanan modeller altında Rasch temelli modeller olan Kısmi Puan Modeli (KPM) ve Dereceleme Ölçeği Modeli (DÖM) ikiden fazla cevaplama kategorisi bulunan maddelerin analizi için geliştirilmiştir. Maddelerin ayırt edicilik parametresi α her madde için aynıdır ve 1' e eşitlenir. Güçlük parametresi olan b parametresi bu modellerde eşik veya adım parametresi olarak adlandırılmaktadır. Eşik parametresi yanıt seçeneklerinin sayısından bir eksiktir (Embretson ve Reise, 2000; Han ve Hambleton, 2014). Örneğin, beş kategorili bir maddenin eşik (b) parametre sayısı dörttür. Bu modellerde bireyin maddeye vereceği yanıt kategoriler birbiriyle karşılaştırılarak kestirilir. Örneğin 0, 1, 2 ve 3 diye puanlanan bir madde için 0 ve 1 alma olasılıkları karşılaştırılır aynı şekilde 1 ve 2, 2 ve 3 için de geçerlidir. İki kategorili maddelerde θ yetenek düzeyindeki bir bireyin i maddesini doğru cevaplama olasılığı diye hesaplama yapılırken bu puanlama çok kategorili bir maddenin puanlamasına genelleştirilirse, yetenek düzeyi θ olan bireyin i kategorili puanlanan bir maddeden x puan alma olasılığı şeklinde hesaplanmaktadır (Han ve Hambleton, 2014). Rasch temelli olmayan modellerden olan Genelleştirilmiş Kısmi Puan Modeli (GKPM) ve Aşamalı Tepki Modeli (ATM)'de ise ayırt edicilik parametresi her madde için farklıdır. KPM ve GKPM arasındaki tek fark her bir madde için ayırt edicilik parametresinin dahil edilmesidir (Han ve Hambleton, 2014). Yani GKPM'de α parametresi 1'e eşitlenince KPM ve GKPM aynı sonuçları verecektir. Aşamalı tepki modeli (ATM) sıralı kategorik yanıtların olduğu durumda kullanılması uygun olan iki parametreli lojistik modelin bir uzantısıdır. ATM'de her bir madde için kategoriler arası eşik (b) parametresi sıralı olmalıdır. Bu parametrenin sıralı olma durumu kısmi kredi ve genelleştirilmiş kısmi kredi modeli için gereklilik değildir (Embretson ve Reise, 2000). Bu durumda bireylerin maddeye vereceği yanıtın ne olacağı iki kategori karşılaştırılarak değil de kategoriler arası birikimli karşılaştırma yaparak kestirilir. Bu birikimli karşılaştırma 0- 1, 2, 3; 0, 1- 2. 3; 0, 1, 2- 3 şeklinde yapılmaktadır. Kategorik puanlanan modellerden olan Sınıflamalı Tepki Modeli (STM) Bock (1972) tarafından ortaya atılmıştır (Embretson ve Reise, 2000). STM'de diğer çok kategorili modellerde olan kategorilerin sıralı bir şekilde olması gerekliliği bulunmaz. Tutum, kişilik ölçeklerinde ve çoktan seçmeli bir testte bu modeller kullanılabilir. Örneğin davranış ölçeğinde yer alan "yemek yemeden önce ellerimi en az beş kez yıkarım" maddesinde yer alan "evet" "belki" "hayır" "belirtmek istemiyorum" yanıt kategorilerini seçebilir ama bu kategoriler sıralı değildir (de Ayala, 2009).

Test ve Madde Bilgi Fonksiyonu

MTK'da bir madde seçme durumunda veya birden fazla testleri karşılaştırma durumunda sıklıkla madde bilgi fonksiyonu ve test bilgi fonksiyonundan yararlanır. MTK yetenek düzeyinde kestirimler yaptığı için bu fonksiyonlar yardımıyla testin veya maddenin hangi yetenek düzeyinde daha iyi kestirimler yaptığını ortaya koymaktadır.

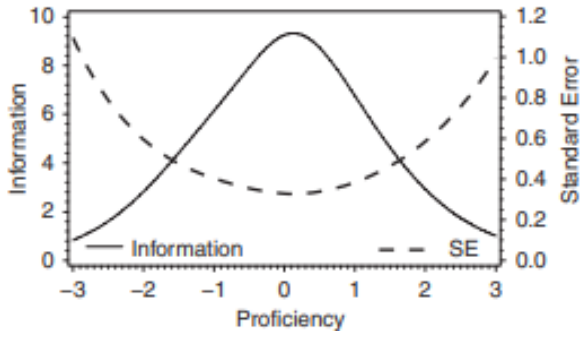
Bilgi fonksiyonları ölçmenin standart hatasını ve güvenirliğini ölçmek için kullanılır. Ölçmenin standart hatası bilginin karekökünün tersi olduğundan dolayı maddenin verdiği bilgi arttıkça hata azalacaktır ve güvenirlikte yükselecektir (deMars, 2010). Madde bilgi fonksiyonları birleşerek test bilgi fonksiyonunu oluşturur (Crocker ve Algina, 2006; de Ayala, 2009) ve maddenin bulunduğu konumuna göre simetrik olup tek bir yetenek düzeyi için maksimum değer alır (de Ayala, 2009). Test bilgi fonksiyonuna ilişkin matematiksel ifade (Eşitlik 4) ve test bilgi fonksiyonu ile ölçmenin standart hatası arasındaki matematiksel ifade (Eşitlik 5) aşağıda verilmiştir (Embretson ve Reise, 2000).

$$I(\theta) = \sum_i^n I_i(\theta) \qquad \text{(Eşitlik 4)}$$

Madde sayısının n olduğu bir test için test bilgi fonksiyonu i. maddeden n. maddeye kadar olan maddelerin bilgi fonksiyonlarının toplamına eşit olup, bu bilgi $I(\theta)$ ile gösterilir.

$$SH(\theta) = \sqrt{\frac{1}{I(\theta)}} \qquad \text{(Eşitlik 5)}$$

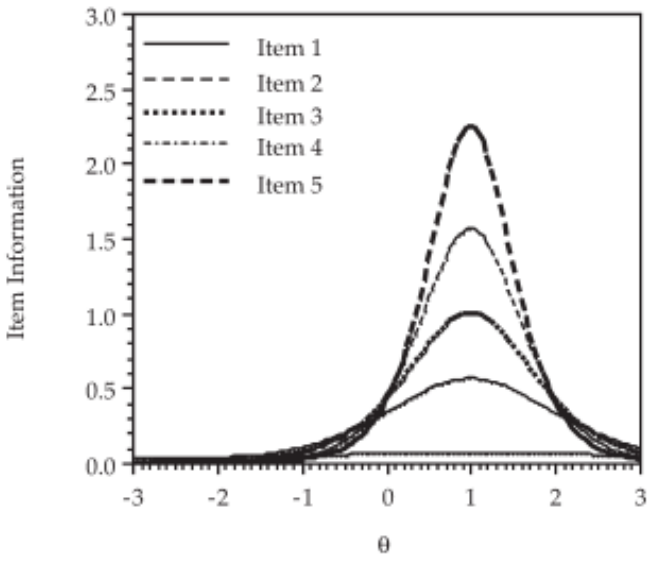
Eşitlik 5'te görüldüğü gibi ölçmenin standart hatası test bilgi fonksiyonu ile ters orantılıdır. Testin sağladığı bilgi arttıkça ölçme sonuçlarındaki standart hata miktarı azalacaktır. Test bilgi fonksiyonu ile ölçmenin standart hatası arasındaki ilişki aşağıda verilen Şekil 4'te detaylı verilmiştir.



Şekil 4. Test Bilgisi ile Yetenek Düzeyi Tahminini Standart Hata Arasındaki İlişki (deMars, 2010)

Şekil 4'te görüldüğü gibi bilgi düzeyinin arttığı yerde standart hata değeri düşmektedir. Test bilgi fonksiyonu yetenek ölçeğindeki bazı yetenek düzeylerinde maksimum bilgi sağlarken bazı düzeyler için daha az bilgi sağlamaktadır. Bu durumda bu testin maksimum bilgi verdiği yetenek düzeyi ve civarındaki bireyler için yetenek kestirimi daha az hatayla yapılmış olur (Embretson ve Reise, 2000; Hambleton vd., 1991).

Madde bilgi fonksiyonu, her yetenek düzeyinde madde karakteristik eğrisinin eğimi ve koşullu varyansa bağlıdır. Eğim ne kadar büyük ve varyans ne kadar küçük olursa, bilgi o kadar büyük olur ve dolayısıyla ölçümün standart hatası o kadar küçük olur. Bu değerlendirme süreciyle standart hatası büyük olan maddeler elenebilir. Madde bilgi fonksiyonları genel olarak çan şeklindedir (Hambleton ve Swaminathan, 1985). Eğim katsayısına bağlı olarak maddenin verdiği bilgi miktarı aşağıda verilen Şekil 5'te gösterilmiştir (de Ayala, 2009).



Şekil 5. Ayırt edicilik parametreleri α1 = 0.5, α2 = 1.5, α3 = 2.0, α4 = 2.5, α5 = 3.0; Güçlük parametresi b = 1.0 olan 5 Maddeye Ait Bilgi Fonksiyonları

Şekil 5 incelendiğinde yetenek düzeyi 1 olan, α parametresi birbirinden farklı olan beş maddeye ait madde bilgi fonksiyonu verilmiştir. Beş madde incelendiğinde ayırt edicilik parametresi daha büyük olanın verdiği bilgi miktarının da daha yüksek olduğu görülmektedir.

En yüksek bilgi, bir ve iki parametreli lojistik modeller için yetenek düzeyinde güçlük parametresinden elde edilir. Bir parametreli model için bilginin maksimum değeri sabittir, ancak iki parametreli modelde maksimum değer, madde ayırıcılık parametresinin

karesiyle doğru orantılıdır yani ayırt edicilik arttıkça bilgi de artmaktadır. Üç parametreli modelde maksimum bilgi aşağıda verilen formül yardımıyla hesaplanmaktadır. Tahmin etme (şans) parametresi ci azaldıkça bilgi artar ve $ci = 0$ olduğunda maksimum bilgi elde edilir. Bu modellere ilişkin test bilgi fonksiyonları Tablo 2'de verilmiştir (Hambleton ve Swaminathan, 1985).

Tablo 2

İki Kategorili Maddelere Ait Madde Bilgi Fonksiyonu

Model	Matematiksel Formül	θ_{max}
1 PL	$I_i(\theta)=P_i\theta_i$	b_i
2 PL	$I_i(\theta)=a^2P_i\theta_i$	b_i
3PL	$I_i(\theta)=\frac{a^2\theta_i}{P_i}[(P_i-ci)^2/(1-ci)^2]$ $[\frac{1}{2}+\frac{1}{2}\sqrt{1+8ci}]$	$b_i+\frac{1}{Dai}\ln$

θ_{max}= θ düzeyindeki maksimum bilgi

bi =*i* maddesinin güçlük parametresi

αi = *i* maddesinin ayırt edicilik parametresi

ci = *i* maddesinin tahmin etme (şans) parametresi

D= 1.702 ölçekleme katsayısı

ln= doğal logaritma fonksiyonu

Madde bilgi fonksiyonu BOBUT uygulamalarında önemli bir rol oynar. BOBUT uygulamalarında ölçüm hassasiyetine maksimum katkı sağlayan maddeler kullanılır. Genel olarak en fazla bilgi sağlayan maddeler bireyin yaklaşık olarak %50 ile %60'lık bir doğru cevap verme olasılığına sahip olduğu maddelerdir (Hambleton vd., 1991).

Bilgisayar Ortamında Bireye Uyarlanmış Testler

Test tarihi boyunca bireysel test uygulaması ve grup test uygulaması arasında her zaman bir kıyas olmuştur. 20. yüzyıl boyunca, tercih neredeyse her zaman grup test uygulaması lehine olmuştur. Toplu olarak uygulanan bir testin karşı karşıya olduğu önemli bir sorun, uygulanacak testin geniş yetenek aralığına hitap etmesi gerektiğidir. Herkesi doğru bir şekilde ölçmek için, testin grubun yetenek ranjına uygun zorluklarda maddeler içermesi gerekir. Örneğin, yetenek düzeyi düşük bireyler için kolay maddeler,

yetenek düzeyi yüksek bireyler için ise zor maddeler olması gerekir aksi halde bireylerin yetenek düzeylerinin birbirinden ayırt edilmesi zorlaşır (Wainer, 2000). Bir testte yer alan maddelerin hem kapsamı temsil edip hem de tüm yetenek ranjına hitap etmesi olanaksız olduğundan toplu test uygulaması yerini bireysel test uygulamalarına bırakması kaçınılmaz olmuştur.

Bireysel olarak uyarlanmış bir testin teorik yapısının ve birçok pratik detayının oluşturulmasında Lord (1970) tarafından yapılan çalışmalar öncülük etmiştir (Wainer vd., 2000). 1980'lerden itibaren ise teknolojinin ilerlemesiyle birlikte ölçme alanında bilgisayar destekli adaptif testlerin (Computerized Adaptive Test- CAT) uygulanması popülerlik kazanmıştır (Rotou vd., 2007). Adaptif testler ülkemizde Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT), Bilgisayarlı Bireyselleştirilmiş Test (BST) gibi çevirilerle adlandırılmaktadır.

BOBUT uygulaması, madde güçlüğü yetenek aralığının orta noktasına denk gelen bir maddeyi sormakla başlar. Madde doğru cevaplanırsa, sorulan bir sonraki madde daha zor olmakla birlikte yanlış cevaplanırsa, bir sonraki madde daha kolay olacak şekilde devam eder. Bu süreç, bireyin yetenek düzeyini belirlemek için önceden belirlenmiş bir sonlandırma kuralı karşılanana kadar devam eder (Wainer vd., 2000).

BOBUT uygulamalarının belirgin bir avantajı, sınavın daha kısa sürede sonlanmasını vadetmesidir. Çünkü sınava giren birey, çok kolay veya çok zor olan sorular ile karşılaşmaz, yetenek düzeyine uygun sorular ile karşılaşır. Sınavı, bireyin yetenek seviyesine uyacak şekilde uyarlamak geleneksel kağıt-kalem testlerinden daha verimli olan adaptif testlere yol açar. Sonuç olarak, BOBUT daha yüksek bir ölçme hassasiyet seviyesine ulaşmak için daha az soru cevaplamayı gerektirir (Rotou vd., 2007; Segall, 2005). BOBUT uygulamalarının, daha kısa test, ölçüm hassasiyeti, talep üzerine test yapabilme ve anında test puanlama ve raporlama sunması BOBUT'u çok çekici kılmaktadır. Ancak, başlangıç maliyetlerinin yüksek olmasının yanı sıra önemli finansal destek ve insan kaynakları gereklidir. Test güvenliği başlangıçta, BOBUT uygulamalarının en büyük avantajlarından biri gibi görünse de sonradan en büyük sorunlarından biri haline gelmiştir. Madde havuzlarının sürekli güncellenmesi, madde ve test güvenliğini sağlamak için gereklidir. Bu durum BOBUT uygulamasını hayata geçirmenin maliyetini büyük ölçüde artırmıştır. BOBUT uygulamalarının sorunsuz olmadığı doğrudur, ancak avantajları dezavantajlarını aşmaktadır (Meijer ve Nering, 1999). Etkili bir BOBUT uygulaması için kullanılacak MTK modeli, madde havuzu, test

başlatma yöntemi, madde seçim yöntemi, yetenek kestirim yöntemi ve sonlandırma kuralına karar vermek önemlidir (Thompson ve Weiss, 2011; Weiss ve Kinsbury, 1984).

Madde havuzu

Farklı psikometrik özelliklere göre maddelerden seçilen en uygun maddelerin oluşturduğu kümeye madde havuzu denir (Reckase, 2010). BOBUT uygulamalarında, madde istatistikleri hesaplanmaz bunun yerine, önceden geliştirilmiş ve istatistikleri hesaplanmış maddeler kullanılır. Eğer bir madde uygun şekilde yapılandırılmamışsa, hiçbir istatistiksel yöntem uygun hale getirmeye yardımcı olamaz (Wainer, 2000). Herhangi bir BOBUT'un başarısı, uygulanan maddelerin seçildiği madde havuzunun kalitesine büyük ölçüde bağlıdır. Havuzdaki maddeler, testi alan bireylerin yetenek düzeyine uygun yeterli bilgi sağlayan özelliklere sahip olmalıdır. Bir başka ifadeyle havuz her yetenek düzeyine uygun güçlükte ve ayırt edicilikte yeterli maddeye sahip olmalıdır (Wise, 1997).

BOBUT uygulamaları, genellikle geniş bir madde havuzundan faydalanır ve bireylere belirli aralıklarla testler uygular. Ancak, bireyin belirli aralıklarla testi alması güvenlik riskleri barındırabilir. Önceki testleri alan bireyler test maddeleri hakkında bilgi edinip, sonraki testleri alacak olan kişilerle bu bilgiyi paylaşabilirler. Aynı zamanda madde havuzu uzun süre kullanıldıktan sonra da testin güvenliği de tehlikeye girebilir (Zhang, 2013). Test güvenliği, geleneksel testlerde olduğu gibi BOBUT uygulamalarında da önemli bir endişe kaynağıdır çünkü test güvenliği doğrudan test geçerliliği ile ilişkilidir (Davey ve Nering, 2002). Bu nedenle, BOBUT uygulamalarında madde kullanım sıklığı ve kapsam dengeleme gibi güvenlik önlemleri alınmalıdır (Reckase, 2010; Zhang, 2013). Mevcut madde seçim yöntemleri genellikle yetenek düzeyi tahmini için en yüksek bilgi değerlerini veren maddeleri seçer. Buradaki amaç yüksek ölçme hassasiyeti sağlamaktır. Ancak, bir madde çok sık kullanılırsa, testi alan bireylerin büyük bir kısmı bu maddeyi görecektir. Bu durum, maddenin maruz kalınma oranını arttırır ve yüksek bir madde maruz kalma oranı ise büyük bir test güvenliği riskine yol açar (Chang ve Ansley, 2006).

Havuz büyüklüğünün artması, bir maddenin daha az sıklıkta gösterilmesini sağlar. Havuzun küçülmesi ise maddenin sık kullanılmasını sağlayarak maddenin aşinalığını arttırır. Bu da maddenin güçlük düzeyinin azalmasına neden olur. Bu durum BOBUT uygulamasının güvenirliğine ve geçerliğine zarar verir çünkü BOBUT'ta amaç madde parametrelerinin değişmeden kalmasını sağlamaktır (Wise, 1997). İki kategorili

puanlanan ve 3 parametreli lojistik model kullanılacak olan bir madde havuzunun en az 100 madde içermesi önerilmiştir (Urry, 1977). Madde havuzu uygun yetenek düzeylerinde bilgi sağlayan kaliteli maddeler kullanılarak geliştirilmişse 100 maddelik bir havuzun bazı amaçlar için hizmet edebileceği önerilmektedir. Ancak, bir havuzun bilgi seviyesinin büyük ölçüde içerdiği madde sayısına bağlı olduğu göz önüne alındığında, 200 veya daha fazla madde içeren bir havuz, çoğu amaç için daha iyi hizmet eder ve çoğu durumda yanlış θ tahmin riskini azaltır (Şahin ve Weiss, 2015).

Test Başlatma Yöntemi

BOBUT uygulamalarında her bireye madde uygulanmadan önce atanan başlangıç θ değeri için birkaç yöntem bulunmaktadır. Birinci yöntem, ortalama güçlük düzeyine karşılık gelen θ=0 olacak şekilde bir madde atamaktır. Ancak, her bireyin θ=0 düzeyinde bir madde alması, bu yetenek düzeyine hitap eden maddelerin çok sık kullanılmasına neden olabileceğinden, bu değer -0.5 ile 0.5 arasında rastgele seçilebilir. İkinci yöntem, bireyin önceki okul başarıları veya test puanlarını kullanarak, yetenek düzeyine uygun bir madde ile testi başlatmaktır. Bu, bireyin önceki performansına dayalı olarak daha doğru bir başlangıç noktası sağlar. Son olarak, bireyin başlangıç yetenek düzeyini tahmin etmek için sosyoekonomik düzey ve motivasyon gibi dış faktörleri kullanarak testi başlatmak da mümkündür (Thompson ve Weiss, 2011).

Madde Seçim Yöntemi

Madde seçimi genellikle madde bilgisi kavramına dayanır; bu kavram, bazı maddelerin bireyin yetenek düzeyi için diğerlerinden daha uygun olduğunu ifade eder. Örneğin, yetenek düzeyi yüksek olan bireye çok kolay bir madde vermek pek mantıklı değildir çünkü birey bu maddeye yüksek olasılıkla doğru cevabı verir. Bu durum ise bireyin yetenek kestirimin hatalı olmasına neden olur (Thompson ve Weiss, 2011). BOBUT uygulamalarında çeşitli madde seçme yöntemleri kullanılmaktadır. Bu araştırmada yaygın kullanılan Maksimum Fisher bilgisi (Maximum Fisher Information-MFB), Kullback-Leibler Bilgisi (Kullback -Leibler Information-KLB), Ağırlıklandırılmış Olabilirlik Bilgi Ölçütü (Likelihood Weight Information Criterion) ve GMAT'in yeni versiyonunda kullanılan ve Han (2012) tarafından önerilen Verimlilik

Dengeli Bilgi (Efficiency Balanced Information- VDB) ile ilgili detaylı açıklamalar aşağıda verilmiştir.

Maksimum Fisher Bilgisi (Maximum Fisher Information- MFB). BOBUT uygulamalarında MFB sıklıkla kullanılan bir yöntemdir (van der Linden ve Pashley, 2000). Bu yöntem, her maddeye verilen cevaptan sonra kestirilen yetenek düzeyi için en yüksek bilgi veren maddenin seçilme sürecini yönetir. MFI yöntemi, her birey için maksimum test bilgisi sağlayan etkili bir madde seçim yöntemi olarak popülerdir. Ancak, BOBUT'un erken aşamalarında (örneğin beş veya daha az madde uygulandığında) ara yetenek kestirimleri (her madde uygulanmasından sonra yapılan yetenek kestirimi) genellikle doğru değildir. Bu nedenle testin başında MFB kriterine göre seçilen maddeler ara yeterlilik tahminlerinde beklenen düzeyde bilgi sağlamaz. MFB yönteminin bir diğer sınırlılığı ise, daha yüksek ayırt edicilik değerlerine sahip maddeleri nispeten daha düşük ayırt edicilik değerlerine sahip maddelerden daha sık kullanmasıdır. Bu dengesiz madde kullanımı madde havuzunun dengeli kullanılması bakımında ciddi sorunlara yol açabilir (Han, 2009).

Kullback-Leibler Bilgisi (Kullbak-Leibler Information- KLB). KLB, istatistiksel tanımı olarak aynı parametre uzayı üzerindeki iki olasılık arasındaki ayrışımı yani asimetrik mesafeyi ölçer. KLB'nin BOBUT bağlamında uygulanması ise ilk olarak Chang ve Ying (1996) tarafından tanıtılmıştır (Wang vd., 2010).

Chang ve Ying (1999) tarafından geliştirilen KLB yöntemi, madde seçiminde MFB'nin kullandığı θ değişimine etki eden "yerel bilginin" aksine "global-küresel bilgiyi" kullanır (Han, 2009). Yerel bilgi θ etrafındaki bilgiyi temsil ederken global bilgi ise bu alanın dışındaki bilgiyi temsil etmektedir. BOBUT uygulamalarında θ konumu hakkında yeterli bilgi olduğunda, yerel bilgi, madde seçimi için bir referans noktası olarak hizmet edebilirken, böyle bir bilginin eksik olduğu durumlarda global bilgi tercih edilebilir. BOBUT esnasında, θ parametre uzayında değiştikçe, maddenin ayırt edicilik gücü hakkında global bir profil oluşturulur ve bu kapsamda sonraki madde seçilmektedir. Chang ve Ying (1999) yaptığı simülasyon çalışmasında, özellikle test uzunluğunun kısa olduğu durumlarda KLB'nin MFB'ye göre bireyin yetenek kestirimini daha az hata ve yanlılık değeri ile kestirdiğini belirtmişlerdir (Chang ve Ying, 1999).

Ağırlıklandırılmış Olabilirlik Bilgi Kriteri (Likelihood Weighted Information Criterion- AOB). Veerkamp ve Berger (1997), BOBUT uygulamalarında madde seçiminde klasik MFB yöntemine alternatif olarak, yetenek kestirimindeki belirsizliği dikkate alan yeni kriterler geliştirmiştir. Bunlardan biri de Ağırlıklandırılmış Olabilirlik Bilgi Kriteri (AOB)' dir. Bu yöntem, yetenek kestiriminin tüm olası değerleri boyunca bilgi fonksiyonunu, verilen yanıtların oluşturduğu olabilirlik fonksiyonuyla ağırlıklandırarak toplar. Test süresince bireyin yeteneği her adımda güncellenmez. Yetenek tahmini yalnızca test sonunda yapılır. Böylece madde seçimi, yetenek kestiriminin belirsizliğini ve olası tüm yetenek seviyelerini dikkate alır. Özellikle testin erken aşamalarında geçici yetenek kestirimlerinin güvensiz olduğu durumlarda, madde seçiminin daha dengeli ve verimli olmasını sağlar. Ayrıca bu yöntem madde havuzunun daha etkili kullanımını destekleyerek, test sürecinde daha güvenilir yetenek kestirimlerine katkıda bulunur.

Verimlilik Dengeli Bilgi (Efficiency Balanced Information- VDB). Verimlilik Dengeli Bilgi (Han, 2012) tarafından önerilen ve GMAT'in Temmuz 2024'te piyasaya sürülen yeni versiyonu olan GMAT Focus Edition'da kullanılan madde seçim yöntemidir.

BOBUT uygulamalarında madde seçim yöntemlerinin çoğu θ noktasında en yüksek bilgiyi veren maddeyi kullanmaya odaklı iken bu yöntemde bilgi belirli bir θ noktasında değerlendirilmez, θ aralığı boyunca değerlendirilir. Madde verimliliği ve bilginin değerlendirilmesi için θ aralığının genişliği, tahminin standart hataları ile belirlenir (Han, 2012; Promono, 2023). Bu yöntemin zayıf noktası, testin başında katılımcının yetenek düzeyini tahmin etmede daha az doğru olmasıdır.

BOBUT uygulaması ilerledikçe kalan maddeler arasındaki bilgi farklılıkları artar bunun sonucunda ise madde bilgisi madde seçiminde daha kritik bir faktör haline gelir. Bu yöntem sayesinde BOBUT uygulamasının başında daha düşük ayırt edicilik değerine sahip maddeler daha olası seçilirken, ilerleyen aşamalarında daha yüksek ayırt edicilik değerine sahip maddeler tercih edilir. Bu yöntem, madde seçim süreci için a-tabakalama yönteminin uyguladığı strateji ile aynıdır, fakat maddeleri tabakalamayı içermez. Ayrıca, a-tabakalama yönteminden farklı olarak her bir madde hakkında c şans parametresi dahil mevcut olan tüm bilgileri dikkate alır (Han, 2012).

Yetenek Kestirim Yöntemi

Bir BOBUT uygulamasında en uygun yetenek kestirim yönteminin seçilmesi çok önemlidir. Çünkü yetenek kestirimi sadece yetenek düzeyinin sonucunu değil, aynı zamanda hangi maddelerin uygulandığını da etkilemektedir. En yaygın kullanılan yetenek kestirim yöntemleri maksimum olabilirlik yöntemi (Maximum Likelihood – MO) ile Owen'ın yöntemi (Owen's method), beklenen sonsal dağılım (expected a posteriori, BSD) ve maksimum sonsal dağılım (maximum a posteriori, MSD) olan Bayesçi yaklaşımlardır (Wang ve Vispoel, 1998).

MO yöntemi birçok istatistiksel yöntemde parametre tahmini için geniş bir şekilde kullanılmaktadır. MO yönteminde sınavı alan birey tüm maddeleri doğru ya da yanlış cevapladığında yetenek kestiriminde sırasıyla pozitif ve negatif sonsuz değer üretir, bu nedenle olabilirlik denkleminin çözümünde sorunlar ortaya çıkmaktadır. Bu sebeple MO yöntemi kullanılacaksa sınavı alan bireyin en az bir maddeyi doğru ve yanlış yanıtlayana kadar Bayes temelli bir yetenek kestirim yöntemi kullanılabilir (Wang ve Vispoel, 1998). Ayrıca bu yöntemde testte yer alan maddelerinin sayısı az olduğunda da sorunlar oluşturur. Bayes yöntemleri, maksimum olabilirlik yöntemleriyle karşılaşılan sorunları aşar (Hambleton vd., 1991), ancak uygun olmayan önsel dağılımlar seçildiğinde yetenek tahminlerinde yanlı kestirimler yapabilir (Lord, 1986). Bayesian yöntemlerinden olan BSD kestirim yöntemi, bireyin yetenek düzeyini tahmin etmek için bireyin yanıt örüntüleri ile ilişkili olabilirlik bilgisi ve popülasyondaki yetenek dağılımının yani önsel bilginin bir kombinasyonunu kullanır. Bu yöntemde bireyin yetenek kestirimi tüm maddelere doğru veya yanlış cevap verildiğinde de yapılır ayrıca MO'ya göre nispeten daha kısa testlerde de kararlı sonuçlar vermektedir (Thissen ve Orlando, 2001). MSD kestirim yöntemi ise BSD'den farklı olarak sonsal dağılımın ortalaması yerine modunu kullanarak bu olasılık dağılımında en yüksek olasılığa sahip yetenek seviyesini (θ) tahmin eder. Bundan sebeple Bayesian yetenek kestirim yöntemlerinde kullanılan önsel dağılımın doğru seçimi, yetenek kestirim sonuçlarının doğruluğu ve güvenilirliği açısından kritik bir rol oynar.

Sonuç olarak MO yöntemi bireyin cevap örüntüsünü temel alarak, MTK varsayımları sağlandığında ve test yeterince uzun olduğunda yansız kestirim yaparken tüm maddeler doğru veya yanlış yanıtlandığında sonlu bir yetenek kestirimi yapamamaktadır (de Ayala, 2009). İyi yapılandırılmış bir madde havuzu ile MO yönteminin nispeten yanlı olmayan ancak nispeten büyük standart hata (SE) ile kestirim

yaptığı, Bayes yöntemlerinin ise önsel ortalamaya doğru nispeten yanlı olduğu ve Bayes yöntemleri arasında BSD'nin nispeten küçük yanlılık ve SE'ye sahip olduğu yönündedir. Bunun yanı sıra BSD, MO ve MSD'ye göre hesaplama açısından daha basit olma avantajına sahiptir. Bayes yöntemleri kestirim için hem mevcut cevap örüntüsünü hem de ön bilgiyi kullanırken, MO sadece cevap örüntüsünü kullanmaktadır (Wang, 1997).

Sonlandırma Kuralları

BOBUT uygulamaları her bireyin aynı sayıda madde aldığı sabit uzunlukta veya değişen uzunlukta tasarlanabilmektedir. Değişen uzunlukta testlerde, madde sayısı bireyin testte gösterdiği performansına göre uyarlanır. Bu tür testleri uygulamak için çeşitli yöntemler vardır. Bunlar (Thampson ve Weiss, 2011; Wainer, 2000);

θ Tahminine Dayalı Sonlandırma. Katılımcının yetenek tahmini (θ) belirli bir madde sonrasında küçük bir miktardan fazla değişmiyorsa test sonlandırılır.

Ölçüm Standart Hatasına Dayalı Sonlandırma. Test, ölçüm standart hatası belirli bir seviyenin altına düştüğünde sona erer.

Madde Havuzuna Dayalı Sonlandırma. Eğer madde havuzunda belirli bir bilgi seviyesini sağlayan maddeler kalmadıysa test sonlanır.

Sabit uzunluk durdurma kuralı altında, BOBUT uygulaması belirli bir sayıda soru uygulandıktan sonra sonlandırılır. Bu durumda, ölçüm hassasiyetinin testin sonunda ne kadar yüksek veya düşük olduğu dikkate alınmaksızın, tüm katılımcılara aynı sayıda soru sorulur. Sabit uzunluk durdurma kuralının en büyük avantajı uygulanmasının oldukça basit olmasıdır. Ancak, bu yöntemin bazı önemli dezavantajları da vardır. Öncelikle, sabit uzunluk durdurma yönteminin uygulanması sonucunda katılımcıların yetenek düzeyi birbirinden farklı derecelerde hassasiyetle ölçülür. Bu durumda bazı katılımcıların yetenek düzeyi daha hassas bir şekilde ölçülürken, diğerlerinin ölçümleri daha az hassas olabilir. Özellikle yetenek düzeyleri çok yüksek veya çok düşük olan katılımcıların yetenek düzeyi genellikle daha büyük ölçüm hataları ile ölçülür. Bu durum testin sonundaki ölçüm sonuçlarının güvenilirliğini olumsuz etkileyebilir. Ayrıca sabit uzunluk durdurma yöntemi BOBUT'un verimliliğini de sınırlayabilir. Bunun sonucunda ise katılımcı yetenek düzeyi hakkında çok az bilgi sağlayan maddeleri yanıtlayabilir bu da

testin gereksiz yere uzamasına ve katılımcılara ek yük getirmesine neden olur (Choi vd., 2011). Sonuç olarak, sabit uzunluk durdurma kuralı basit ve uygulanması kolay olmasına rağmen ölçüm hassasiyetini ve test verimliliğini olumsuz etkileyebilecek dezavantajlara sahiptir. Bu nedenle, adaptif testlerin tasarımında bu dezavantajların göz önünde bulundurulması önemlidir.

BOBUT uygulamalarında testler bireye özgü sınav anında oluşmaktadır. Bu uygulamalarda bireyin yetenek düzeyi hakkında en fazla bilgi veren madde seçildiği için test kapsamını temsil etmeyebilir ve bu durumda kapsam geçerliği zarar görebilir (Shin vd., 2009). Aynı zamanda bireyin yetenek düzeyi hakkında en çok bilgi veren maddelerin sıklıkla kullanılması maddelerin ifşa olmasına bu durum ise testin güvenirliğinin düşmesine neden olabilir (Georgiadou vd., 2007). Bu dezavantajlar için kapsam dengeleme ve madde kullanım sıklığı yöntemleri geliştirilmiştir. Bu araştırmada belirtilen dezavantajları minimum düzeye indirmek için kapsam dengeleme ve madde kullanım sıklığı yöntemleri kullanılmıştır.

Kapsam Dengeleme

BOBUT uygulanmasında karşılaşılan sorunlardan biri bireylerin BOBUT sürecinde aldığı maddelerin testin kapsamını karşılayıp karşılayamadığıdır (Kingsbury ve Hauser, 2005; Kingsbury ve Zara, 1989). BOBUT uygulamasında bir sonraki maddelerin seçimi, bireyin yetenek düzeyine hakkında en fazla bilgiyi veren maddelerden seçildiği için bazen testin kapsamı dengelenmeyebilir (Shin vd., 2009). Bu uygulamalarda madde seçerken maddenin sadece ayırt ediciliği değil aynı zamanda test geçerliğini ve güvenirliğini etkileyen kapsam dengelemeye (content balancing) ve madde kullanım sıklığına (item exposure) da bakmak gerekmektedir (Gu ve Reckase, 2007). Bunun için çeşitli kapsam dengeleme yöntemleri geliştirilmiş ve incelenmiştir. BOBUT'ta kapsam dengelenesini sağlamak için üç farklı yaklaşım önerilmiştir. Bunlar arasında Kingsbury ve Zara (1989) tarafından geliştirilen kısıtlanmış BOBUT (Constrained CAT), Stocking ve Swanson, (1993) tarafından geliştirilen ağırlıklı sapmalar modeli (Weighted Deviations Model-WDM), van der Linden (2000) ve van der Linden ve Reese (1998) tarafından geliştirilen gölge test yaklaşımı bulunmaktadır. Kapsam dengeleme yöntemleri arasındaki en önemli fark temel felsefelerinden kaynaklanır. Her yöntem madde seçiminin hangi esasa dayanması gerektiği, testin doğasına ilişkin farklı görüşü temsil eder (van der Linden, 2005). Kapsam dengesi sağlanmış bir BOBUT, bireylere testin

içerik alanlarının her birini yeterince temsil eden bir test sağlar (Kingsbury ve Zara, 1989).

BOBUT uygulamalarında kullanılan popüler kapsam dengeleme yöntemlerinden biri Kingsbury ve Zara (1989) tarafından geliştirilen kısıtlanmış BOBUT yöntemidir. Bu yöntemde her konu alanı için önce hedef madde yüzdesi belirlenmektedir. Hedef madde yüzdesi kapsama uygun olacak şekilde yapılmaktadır. BOBUT uygulamasında maddeler bireyin yanıtlaması için geldikçe, konu alanı ve hedef alanındaki maddeler için gerçek ve hedef yüzde değerleri hesaplanmaktadır. Bir sonraki madde seçimi gerçek ve hedef değerleri arasındaki farkın en büyük olduğu konu alanındaki maddeden seçilmektedir. Bu yöntem, test içeriğini madde sayısı yerine yüzde olarak değerlendirdiği için hem sabit hem de değişen uzunluklu BOBUT için kullanılabilmektedir (Han, 2018). Bundan dolayı bu araştırmada kapsam dengeleme yöntemi olarak kısıtlanmış BOBUT yöntemi kullanılmıştır.

Madde Kullanım Sıklığı Kontrolü

Kağıt-kalem testlerinde kullanılan maddeler bir kez kullanılmak üzere tasarlanırken, BOBUT uygulamalarında kullanılan maddeler zaman içinde tekrar tekrar kullanılmaktadır. Ancak bazı maddeler daha çok kullanılır. Bu durumda bazı grupların bu maddelere cevap verme davranışı değişir. Birey daha önce yanıt verdiği bir maddeyle tekrar karşılaştığında problem çözme becerisi dayanarak cevap vermekten çok ezberlenmiş bilgiyi kullanır (Han, 2018). BOBUT uygulamalarında bireyin yetenek düzeyi hakkında en çok bilgi veren maddelerin sıklıkla kullanılması bireylerde madde aşinalığına neden olur. Bu durumda maddenin güçlük düzeyinde bir azalmaya ve bireyin yetenek düzeyinin olduğundan yüksek hesaplanmasına neden olur (Georgiadou vd., 2007). Maddenin sık kullanması testin adil ve geçerli olması açısından ciddi tehlike oluşturabilir bu durum test güvenliğine zarar verir (Georgiadou vd., 2007; Han, 2018; van der Linden, 2000). Bu durumda maddenin kullanım sıklığı (item exposure) kontrolü gerekmektedir.

Madde kullanım sıklığını kontrol etmek için çeşitli yöntemler geliştirilmiştir. Bu yöntemler, bazı maddelerin aşırı seçilmesini önlemeyi ve nadiren veya hiç seçilmeyen maddelerin kullanım oranını artırmayı amaçlar (Revuelta ve Ponsoda, 1998). Madde kullanım sıklığı tesadüfi seçme (randomization), koşullu seçme (conditional selection) ve

tabakalı (stratification) olmak üzere tartışılan üç farklı stratejidir (Davis 2002; Davis ve Dodd, 2003).

Tesadüfi seçme stratejisi, belirli bir yetenek seviyesinde maksimum bilgi veren madde yerine, o yetenek seviyesine yakın bir grup maddeden rastgele bir maddeyi seçer. 5-4-3-2-1 yöntemi (McBride ve Martin, 1983) ve Randomesque yöntemi (Kingsbury ve Zara, 1989) yaygın olarak kullanılan iki tesadüfi seçme stratejisidir. Bu yöntemlerin uygulanması çok basittir ancak madde kullanım sıklığını belirli bir yetenek düzeyine sınırlanmasını garanti etmezler (Davis, 2002; Keng, 2008). Tabakalı seçme stratejisi Chang ve Ying (1999) tarafından öne sürülen bir stratejidir. Tabakalı seçme stratejisi temelinde geliştirilen yöntemler madde havuzunda yer alan maddeleri ayırt edicilik parametrelerine göre k farklı tabakaya ayırır. Maddelerin ayırt edicilikleri birbirinden çok farklıysa tabaka sayısı artar ama ayırt edicilikleri benzerse daha az tabaka kullanılır. Bu yöntemlerde testin başında en düşük ayırt ediciliğe sahip tabakadan maddeler seçilir, testin sonunda ise en yüksek ayırt ediciliğe sahip tabakadan maddeler seçilir. Bu sayede düşük ve yüksek ayırt ediciliğe sahip maddelerin seçilme olasılığını eşitleyerek madde kullanım sıklığını dengelenmektedir (Chang ve Ying, 1999; Davis, 2002). Koşullu seçme ise, bir maddenin uygulanma olasılığının belirli bir kritere bağlı olarak kontrol edildiği stratejilerdir. Sympson-Hetter (SH), koşullu Sympson-Hetter (conditional Sympson-Hetter), Stocking-Lewis (Stocking and Lewis multinomial procedures) ve Davey-Parshall (Davey-Parshall procedures) kullanılan koşullu seçme yöntemlerindendir (Davis, 2002). Bu yöntemler arasında Sympson-Hetter (SH) yöntemi (Sympson ve Hetter, 1985) en yaygın kullanılan koşullu seçme yöntemidir (Wang, 2017).

Benzer yetenek düzeylerine sahip katılımcılarına aynı maddelerin tekrar tekrar sunulmasını engelleyerek madde havuzunun daha dengeli kullanılmasına katkı sağlamak (Han, 2012) amacıyla bu araştırmada BOBUT simülasyonlarında madde kullanım sıklığını kontrol etmek için Kingsbury ve Zara (1989) tarafından geliştirilen Randomesque yöntemi kullanılmıştır. Reckase (2003) çalışmasında BOBUT uygulamalarında, içerik dengeleme ve "randomesque" gibi madde kullanım sıklığı kontrolünün bilgisayar ortamında bireye uyarlanmış testlerde kullanılan iki önemli prosedür olduğunu bu bakımından gerçek/canlı BOBUT koşullarını daha gerçekçi biçimde temsil ettiğini ifade etmektedir.

İlgili Araştırmalar

Bilgisayar ortamında bireye uyarlanmış testler üzerine yapılan araştırmalar yurtdışında yoğun olmakla birlikte Türkiye'de de bu konu üzerinde yapılan araştırmalar gün geçtikçe artmaktadır. Bu bölümde yüksek riskli sınavlarla ilgili yapılan BOBUT araştırmalarına, kağıt-kalem testi ile BOBUT uygulamasının karşılaştırıldığı araştırmalara ve ikili puanlanan maddeler ile ilgili araştırmalara yer verilmiştir. Öncelikli olarak yurt içinde daha sonra ise yurt dışında yapılan çalışmaların özetlerine kronolojik sırayla yer verilmiştir.

Kaptan (1993) çalışmasında bireyin yeteneğin kestirilmesinde kâğıt-kalem testi ile BOBUT yöntemini karşılaştırmıştır. Çalışmada Rasch modeli, MO yetenek kestirim yöntemi, çok aşamalı madde seçim yöntemi kullanmıştır. Araştırma sonuçlarına göre; BOBUT uygulaması %55 oranında zamandan tasarruf sağlamıştır. Ayrıca %70 tasarrufla daha az maddeyle benzer yetenek kestirimi yapmıştır.

Kalender (2011) Öğrenci Seçme Sınavı (ÖSS) fen alt testinin kâğıt-kalem formundan elde edilen yetenek kestirimi ile ÖSS'nin BOBUT uygulamasından elde ettiği yetenek kestirim sonuçlarını karşılaştırmıştır. Araştırmada 3PL MTK modeli, MO ve BSD yetenek kestirim yöntemi, MFB madde seçim yöntemi ve standart hata eşik değeri ile sabit madde sayısı test sonlandırma kullanılmıştır. Araştırmanın bulgularına göre BSD yetenek kestirim yöntemi MO'ya göre daha iyi kestirimler yaparken aynı zamanda daha fazla madde kullanmıştır. BOBUT uygulaması ile kağıt-kalem formatındaki ÖSS yetenek kestirimleri arasında 0.74 düzeyinde korelasyon bulunmuştur. Ayrıca BOBUT uygulamasında kullanılan soru sayısı ortalama 18.4 olmuştur. Bu sonuç ile bireylere uygulanan madde sayısında kağıt-kalem testine kıyasla yaklaşık %50 oranında düşüş sağlanmıştır.

Bulut ve Kan (2012) yaptığı çalışmada ALES'in BOBUT olarak uygulanabilirliğini incelemişlerdir. Araştırmada örneklem büyüklüğü 1000 kişi olan 2008 yılı ALES verisi kullanılmıştır. 3PL'ye göre madde parametreleri kestirilmiştir. Yetenek kestirim yöntemi olarak BSD, madde seçme yöntemi olarak MFB, sonlandırma kuralı olarak da değişen uzunluk (.25, .30, ve .40 standart hata) kullanılmıştır. BOBUT uygulamaları, öğrencilere uygulanan madde sayısını %44 ile %88 arasında önemli ölçüde azaltmıştır. BOBUT uygulaması ile kâğıt kalem testi sonuçları arasında 0.93 ve üzerinde korelasyonlar elde edilmiş ve BOBUT uygulamasının güvenilir sonuçlar verdiği sonucuna ulaşılmıştır.

Sayman Ayhan (2015) ÖSS'nin kağıt-kalem formunun BOBUT olarak uygulanabilirliğini incelemiştir. Bunun için ÖSS'nin 2005 yılı örneklem büyüklüğü 2244 ve madde sayısı ise 45 olan fen alt testi verileri araştırma kapsamında incelenmiştir. Araştırmada yetenek kestirim yöntemi olarak MO ve BSD, sonlandırma kuralı olarak da sabit (10, 15 ve 25 madde) ve değişen uzunluk (.10, .20 ve .30 standart hata) yöntemi kullanmıştır. Araştırmanın bulgularına göre, ÖSS kağıt-kalem testi ile BOBUT uygulaması arasında anlamlı ilişki olduğu yetenek kestirim yöntemi olarak BSD'nin tüm koşullarda MO'ya göre daha iyi olduğu, testin sonlanması için BSD'nin MO'ya göre daha fazla madde gerektirdiği ve test sonlandırma kuralı olarak 0.30 standart hatanın en iyi sonuçlar verdiği sonucuna ulaşılmıştır. ÖSS'nin BOBUT uygulamasının kağıt-kalem formuna alternatif olabileceği önerilmiştir.

Şenel (2017) görme engelli öğrencilerin bilgisayar ortamında bireye uyarlanmış testte (BOBUT) ve okuyucu-kodlayıcı yardımıyla aldıkları kâğıt-kalem testlerindeki dinlediğini anlama becerilerini karşılaştırmıştır. Araştırmada 3PL MTK modeli, MO yetenek kestirim yöntemi, MFB madde seçim yöntemi, kestirilen son iki yetenek düzeyine ilişkin standart hatalar arasındaki fark ve standart hata değeri 0.30 olmak üzere iki farklı sonlandırma kuralı kullanılmıştır. BOBUT uygulamasında herhangi bir madde kullanım sıklığı yöntemi ve kapsam dengeleme yöntemi kullanılmamıştır. Araştırma sonuçlarına göre, öğrencilerin BOBUT uygulamasında aldıkları yetenek puanları, okuyucu-kodlayıcı yardımıyla aldıkları kâğıt-kalem test puanlarından anlamlı derecede düşük olmuştur. Ancak, iki test türü arasındaki yetenek puanları arasında pozitif yönde yüksek bir ilişki tespit edilmiştir.

Gür ve Karabay (2018) Motorlu Taşıt Sürücü Kursiyerleri Sınavı'nın BOBUT uygulaması ile kâğıt kalem testinin simülatif olarak karşılaştırdıkları araştırmada 5000 kişiden oluşan gerçek veriye dayalı bir örneklem ve 240 maddeden oluşan bir madde havuzu kullanmışlardır. Araştırmada yer alan maddelerin parametre kestiriminde 2PL model, yetenek kestirimi için MSD, testin sonlandırılması için ise .30 standart hata değeri kullanılmıştır. Motorlu Taşıt Sürücü Kursiyerleri Sınavı'nın kağıt-kalem testi 50 madde ile sonlanırken, BOBUT uygulamasında adaylara minimum 27, maksimum 50 ve ortalama 32.52 madde uygulanmıştır. Bu iki sınav uygulamasında elde edilen yetenek düzeyleri arasında .94 düzeyinde korelasyon bulunmuştur.

Çıkrıkçı vd., (2020) yaptıkları araştırmada Türkiye'de sürücü adayları için yapılan ehliyet sınavın BOBUT olarak uygulanabilirliğini test etmişlerdir. Araştırma kapsamında 913 çoktan seçmeli madde geliştirilmiş gerekli düzeltmelerden sonra madde havuzu

toplamda 892 madde ile tamamlanmıştır. Araştırmada 3PL MTK modeli, MFB madde seçim yöntemi, BSD ve MO yetenek kestirim yöntemi kullanılmıştır. Test sonlandırma yöntemi olarak değişen uzunluk (0.30 ve 0.40 standart hata) ve sabit uzunluk (50 madde) kullanılmıştır. Araştırmada ayrıca canlı BOBUT uygulaması için post hoc simülasyonları yapılmıştır. Canlı BOBUT uygulaması 628 maddelik bir madde havuzu ve 280 kişiden oluşan bir örneklem üzerinden yürütülmüştür. Araştırma sonucunda BSD'nin daha güvenilir sonuçlar verdiği, katılımcılara uygulanan test süresinin BOBUT uygulamasıyla 15 dakikaya düştüğü, sabit soru sayısına sahip BOBUT uygulamasının birey yeteneğini güvenilir bir şekilde ölçebildiği sonucuna ulaşılmıştır.

Çoban (2020) yaptığı araştırmada, YÖKDİL'in bilgisayar ortamında bireye uyarlanmış test, çok aşamalı bireyselleştirilmiş test (ÇABOBUT) ve bilgisayarlı sınıflama testi (BST) bireyselleştirilmiş test yaklaşımlarından hangisinin daha uygun olduğunu belirlemeye çalışmıştır. BOBUT uygulamasında yetenek kestirim yöntemi olarak BSD, madde seçim yöntemi olarak MFB ve sonlandırma kuralı olarak sabit uzunluk yöntemi kullanılmıştır. Ayrıca araştırmada kısıtlanmış BOBUT kapsam dengeleme yöntemi ile Randomesque madde kullanım sıklığı yöntemi kullanılmıştır. Araştırmada kullanılan madde havuzu ve birey yetenek parametreleri simülasyon ortamında üretilmiştir. Sonuç olarak BOBUT'un ÇABOBUT ve BST'ye göre ölçme kesinliği açısından daha iyi olduğu sonucuna ulaşmıştır.

Kaya (2022) yaptığı araştırmada bir İngilizce yeterlilik okuma alt testinin kağıt-kalem formu ile BOBUT formundan elde edilen puanları çeşitli psikometrik özellikler açısından karşılaştırmıştır. Araştırmada 1182 kişiden ve 35 çoktan seçmeli maddeden oluşan veri setinde gerçek veriye dayalı post hoc simülasyonlar yapılmıştır. Araştırmada yetenek kestirim yöntemi olarak BSD, madde seçim yöntemi olarak MFB, test sonlandırma kuralı olarak sabit ve değişen uzunluklu sonlandırma yöntemleri kullanılmıştır. Araştırmanın sonucunda İngilizce yeterlik sınavının kağıt-kalem formu ile BOBUT formunun benzer yetenek kestirimi yaptığı sonucuna ulaşılmıştır. Ayrıca simülasyon sonuçları, testin BOBUT versiyonu 0.50 ve 0.40 standart hata eşik değerlerinde veya sabit uzunluk testi bitirme kuralı altında 15, 20 ve 25 maddede sonlandırıldığında, testin kağıt-kalem formuna kıyasla önemli ölçüde daha az sayıda maddenin uygulanabileceğini göstermiştir.

Yurt dışında yapılan çalışmalar kronolojik sırada incelendiğinde de Beer ve Visser (1998) tarafından Güney Afrikalı lise öğrencilerine uyguladıkları Genel Akademik Yetenek Testi'nin (GSAT), kağıt-kalem, BOBUT ve bilgisayar destekli doğrusal test

(CBT) formlarının karşılaştırılması amaçlanmıştır. Araştırmada madde parametrelerini kestirmek için 3PL MTK modeli kullanılmıştır. Çalışmanın genel amacı, BOBUT sonuçları ile kâğıt kalem sonuçlarını karşılaştırıp, BOBUT uygulamalarını iyileştirmektir. Hem BOBUT hem de kâğıt kalem testi 242 öğrenciye uygulanmıştır. Araştırmanın bulgularına göre öğrencilerin kâğıt kalem testi puanları CBT ve BOBUT uygulamalarındaki puanlardan daha yüksek olduğu sonucuna ulaşılmıştır. Bu sonuçların farklı olması öğrencilerin bilgisayar ortamındaki sınav uygulamalarına aşina olmadığı şeklinde raporlanmıştır. Ayrıca GSAT BOBUT uygulaması yaklaşık olarak 30 dakika sürmüş olup kağıt-kalem uygulamasına kıyasla %75 zaman tasarrufu sağlamıştır.

Costa vd., (2009) Brasilia Üniversitesi'nin İngilizce testinden 246 madde içeren bir madde havuzu üzerinden beş farklı simülasyon çalışması gerçekleştirerek MFB, KLB ve maksimum beklenen bilgi (Maximum Expected Information-MBB) madde seçim yöntemlerinin performansını değerlendirmişlerdir. Araştırmada 3PL MTK modeli ile BSD yetenek kestirim yöntemi ile 0.4 ve 0.2 standart hata değişen uzunlukta sonlandırma yöntemi ve sabit uzunlukta sonlandırma kuralı kullanılmıştır. Araştırmanın sonucunda 0.4 standart hata sonlandırma kuralı uygulandığında MFB yöntemi için kullanılan ortalama madde sayısının 23 olduğu, KLB için 22 ve MBB için 21 olduğu gözlemlenmiştir. 0.2 standart hata sonlandırma kuralı uygulandığında ise, kullanılan madde sayısının yaklaşık altı kat arttığı (MFB için 136, KLB için 137 ve MBB için 138) sonucuna ulaşılmıştır. Analiz edilen madde havuzu ile incelenen üç madde seçim yöntemi arasında θ tahmini açısından belirgin bir istatistiksel fark olmadığı sonucuna ulaşılmıştır. Sabit uzunlukta (25 madde) test sonlandırma kuralı ile yapılan BOBUT uygulamasında tüm madde seçim yöntemlerinde θ kestirim sonuçları, oldukça iyi bir şekilde gerçekleştirilmiştir. Kestirimler ile gerçek değerler arasındaki en düşük korelasyon (0.98) MFB yöntemi ile elde edilmiştir. Ayrıca araştırmada θ=-1.50, θ=0 ve θ=1.50 olmak üzere üç farklı başlangıç θ değerleri ile sonlandırma kuralı 25 madde olacak şekilde ayarlanarak adayların θ kestirimlerinde madde seçiminin nasıl etkilendiği değerlendirilmiştir. θ=1.50 olduğunda, üç madde seçme yönteminin benzer performans gösterdiği ve standart hataların %50'sinin 0.28'den küçük veya eşit olduğu; θ=0 tahmininin standart hata değerlerinin üç madde seçme yöntemi için 0.24 ile 0.30 arasında değiştiği ve θ=−1.50 olduğunda standart hataların ortalama olarak 0.25 ile 0.32 arasında değiştiği sonucuna ulaşılmıştır.

Han (2009) yaptığı araştırmada, BOBUT uygulamalarında beş farklı madde seçme yöntemlerinin etkililiğini simülasyon ortamında karşılaştırmıştır. İlk yöntemde,

madde seçme algoritması, geçici θ'ye en yakın olan beş maddeden birini rastgele seçmektedir. İkinci yöntem olarak MFB, üçüncü yöntemde de MFB yaklaşımı kullanılmıştır ancak madde seçim algoritması farklı bir madde kullanım sıklığı mekanizması kurulmuştur. Dördüncü yöntem olarak kademeli maksimum bilgi oranı (gradual maximum information ratio- GMIR) ve beşinci yöntem olarak da madde kullanım sıklığı kullanılan GMIR olmuştur. Araştırmada 10.000 kişiden oluşan bir örneklem ve GMAT sınavından elde edilen 500 çoktan seçmeli test maddesinden oluşan bir madde havuzu kullanılmıştır. Araştırmada BSD yetenek kestirim yöntemi kullanılmıştır. Araştırmanın sonucunda birinci, üçüncü ve dördüncü madde seçim yöntemi ile hiçbir maddenin madde kullanım sıklığı sınırına (0.20) kadar kullanılmadığı ve madde kullanım sıklığı oranlarının dengeli olduğu sonucuna ulaşılmıştır. Öte yandan ikinci yöntemin son derece dengesiz bir madde kullanımına yol açtığı belirtilmiştir. Son olarak beşinci yöntemin hem madde kullanım sıklığı sınırına kadar kullanılan hem de hiç kullanılmayan maddenin olmadığı bir başka ifadeyle madde havuzunun çok iyi dengelendiği sonucuna ulaşılmıştır. Araştırmanın sonucu olarak GMIR yönteminin, test hassasiyetinin ödün verilmesini en aza indirirken, MFB yöntemi ile karşılaştırıldığında madde havuzu kullanımını büyük ölçüde artırdığı sonucuna ulaşılmıştır.

Zitny vd., (2012) yaptığı çalışmada 29 maddeden oluşan Zeka Potansiyeli (TIP) ve 24 maddeden oluşan Viyana Matris (VMT) bilişsel yetenek testinin kağıt-kalem ve BOBUT uygulamasının sonuçlarını gerçek veriye dayalı olarak post hoc simülasyonu ile karşılaştırmıştır. Araştırmanın örneklemini 803 lise öğrencisi oluşturmaktadır. Araştırmada madde parametre kestirimi için 3PL MTK modeli, yetenek kestirim yöntemi olarak BSD, sonlandırma yöntemi olarak da değişen uzunluk (.50 standart hata) kullanılmıştır. Araştırmanın sonucunda TIP BOBUT uygulaması kağıt-kalem uygulamasına göre %55, VMT BOBUT uygulaması ise kağıt-kalem versiyonuna göre %54 madde tasarrufu sağlamıştır. BOBUT uygulamasında öğrencilerin %4.7'si madde havuzundaki tüm maddeleri yanıtlamıştır. Ayrıca testlerin kağıt-kalem ve BOBUT uygulamasında öğrencilerin yetenek kestirim sonuçlarının benzer olduğu sonucuna ulaşılmıştır.

Wang vd., (2012) yaptığı araştırmada Çin Yeterlilik Testinin (CPT) BOBUT olarak uygulanabilirliğini incelemişlerdir. Çalışmada A1 ve A2 seviyeleri için dinleme ve okuma bölümlerini içeren BOBUT uygulaması geliştirilmiştir. Araştırmada A1 seviyesinde toplam 830 katılımcı, A2 seviyesinde ise toplam 746 katılımcı yer almıştır. Her katılımcı A1 ya da A2 seviyesinde hem dinleme hem de okuma bölümlerini

tamamlamaktadır. Her seviyedeki her testin süresi 30 dakika olup, testin uzunluğu 35 maddedir. Katılımcı, aynı seviyede iki teste de katılmak zorunda olduğu için her katılımcı için toplam test süresi 60 dakikadır. Çalışmada madde parametreleri 3PL MTK modeli, maksimum bilgi (maximum information) madde seçim yöntemi, MO, BSD, MSD yetenek kestirim yöntemleri kullanılmıştır. Yetenek kestirim yöntemlerinin doğruluklarını karşılaştırmak için RMSD (Root Mean Square Deviation) kullanılmıştır. Araştırmanın sonucunda BSD yetenek kestirim yönteminin, MO ve MSD ile karşılaştırıldığında genel olarak daha düşük bir RMSD ile sonuçlandığı görülmüştür. Araştırmacılar BOBUT uygulamalarında yetenek kestirim yöntemi olarak BSD'nin kullanılmasını önermiştir.

Hogan (2013) yaptığı araştırmada devlet tarafından ortaokul düzeyinde yapılan yüksek riskli sınavların BOBUT uygulaması, bilgisayarlı sınıflama testi (BST) ve kağıt-kalem testinin sonuçlarını karşılaştırmıştır. Araştırmanın amacı BOBUT ve BST'nin etkililik ve hassasiyet derecesini kağıt-kalem testlerine kıyasla araştırmak ve simülasyon sonuçlarının okul test koordinatörlerinin yüksek riskli bilgisayar tabanlı testler hakkındaki algılarını nasıl değiştirdiğini keşfetmektir. Araştırmada MO yetenek kestirim yöntemi, MFB madde seçim yöntemi ve hem sabit hem de değişen uzunlukta test sonlandırma yöntemi kullanılmıştır. Araştırmanın bulgularına göre BOBUT uygulaması minimum beş maksimum 24 madde ile sonlandığı, test koordinatörlerinin bilgisayarlı testler hakkında süreçte deneyim kazandığı sonucuna ulaşılmıştır. Ayrıca yüksek riskli sınavların BOBUT olarak uygulanabilirliğini incelemek için öncelikle simülasyon çalışması yapılması gerektiğini ve sınav koordinatörlerinden geri bildirim alınmasının faydalı olacağı bildirilmiştir.

Ogunjimi vd., (2021), Covid-19 pandemisi sırasında Nijerya'daki yüksek riskli sınavlarda önemli aksamalara uğramasının nedeniyle bu sınavların BOBUT olarak uygulanabilmesi için simülasyon çalışması yapmışlardır. Çalışmanın amacı, sabit ve değişen uzunlukta sonlandırma yöntemlerini bir BOBUT ortamında simüle ederek Nijerya'daki yüksek riskli sınavların ölçme doğruluğu ve madde kullanım sıklığı üzerindeki etkilerini araştırmaktır. Araştırmada 3PL MTK modeli, MFB madde seçim yöntemi, sabit uzunlukta (30 madde) ve değişen uzunlukta (.35 standart hata) sonlandırma yöntemleri, MO yetenek kestirim yöntemi ve Randomesque madde kullanım sıklığı yöntemi kullanılmıştır. Madde havuzu 300 maddeden oluşmakla birlikte örneklem büyüklüğü ise 5000 olarak belirlenmiştir. Araştırmada sabit test uzunluğunun değişen test uzunluğuna göre ölçme hassasiyeti açısından iyi sonuçlar verdiği bulgusuna ulaşılmıştır.

Sabit test sonlandırma yöntemi ölçme hassasiyeti açısından daha iyi sonuçlar vermesine karşın madde kullanım sıklığı değişen uzunlukta test sonlandırma yöntemine göre yüksek bulunmuştur.

Yurt içinde ve yurt dışında BOBUT uygulamaları ile ilgili çok sayıda araştırma yapılmıştır. Yurt içinde yapılan çalışmalar incelendiğinde, çalışmaların çoğunluğunda BOBUT uygulamasının her aşamasının genellikle tek bir yöntem ile yürütüldüğü görülmüştür. Madde parametre kestirimi için 3PL MTK modeli, yetenek kestirim yöntemi olarak BSD ve madde seçim yöntemi olarak ise MFB yöntemi sıklıkla kullanılmıştır.

Yurtdışında yapılan çalışmalar incelendiğinde de çalışmalarda yaygın olarak 3PL MTK modeli, BSD yetenek kestirim yöntemi, MFB madde seçim yöntemi kullanılmıştır. Sonlandırma kuralları olarak sabit uzunluk ve değişen uzunluk (standart hata) yöntemleri kullanılmış olup, bazı araştırmalar sabit uzunlukta testlerin ölçme hassasiyeti açısından daha iyi sonuçlar verdiği bazı araştırmalar ise değişen uzunluktaki testlerin ölçme hassasiyeti açısından daha iyi sonuçlar verdiği sonucuna ulaşmıştır. Ayrıca, BOBUT uygulamalarının kağıt-kalem ve bilgisayarlı doğrusal testlere (CBT) göre daha az zaman aldığı ve daha az sayıda madde ile benzer ölçme doğruluğu sağladığı yaygın bir bulgu olarak öne çıkmıştır. Genel olarak hem yurt içi hem de yurt dışı araştırmalar BOBUT'un ölçme doğruluğu, madde tasarrufu ve uygulama süresi açısından kağıt-kalem testine kıyasla avantajlar sunduğunu göstermektedir.

Yüksek riskli sınavlarla ilgili yapılan BOBUT çalışmalarında, BOBUT'un her aşaması için genellikle tek bir yöntem kullanıldığı görülmüş ve BOBUT'un her aşaması için farklı yöntemlerin karşılaştırıldığı çalışmaya rastlanılmamıştır. Ayrıca yurt içinde yapılan benzer çalışmalarda canlı BOBUT uygulaması yapılan sınırlı (Kalender, 2011) çalışmaya rastlanılmıştır. Bu sebeple bu çalışmada alan yazındaki yüksek riskli sınav çalışan araştırmalardan farklı olarak yetenek kestirim yöntemi olarak MO ve BSD; madde seçim yöntemi olarak MFB, KLB, AOB, VDB; sonlandırma kuralı olarak hem sabit hem de değişen uzunluk yöntemleri 40 farklı koşul altında incelenmiştir. Aynı zamanda benzer araştırmalardan farklı olarak, AOB ve GMAT'in yeni versiyonunda kullanılan VDB madde seçim yöntemi test edilmiştir. Ayrıca yurt dışında yapılan çalışmalarda madde kullanım sıklığı ve kapsam dengeleme yöntemlerinin yurt içinde yapılan çalışmalara göre daha yaygın kullanıldığı ve önerildiği için bu araştırmada bu yöntemlerde kullanılmıştır.

BÖLÜM 3

YÖNTEM

Bu bölümde araştırmanın modeli, çalışma grubu, verilerin toplanması ve verilerin analizi aşamaları hakkında bilgi verilmiştir.

Araştırmanın Modeli

Bu araştırma iki aşamadan oluşmaktadır. İlk aşamada ALES kağıt-kalem testinden elde edilen verilerden MTK'ya dayalı madde parametre kestirimi ile başlanmıştır. Devamında kestirilen madde parametreleri kullanılarak farklı madde seçim, yetenek kestirimi ve test sonlandırma yöntemlerinin test edildiği simülasyon ortamı oluşturulmuştur. Bu amaçla araştırmanın ilk aşaması hali hazırda var olan ALES'in farklı simülasyon koşulları altında en uygun ALES-BOBUT uygulamasının hangi madde seçim, yetenek kestirim ve test sonlandırma yöntemlerinin karşılaştırılması sonucunda elde edildiğinin araştırıldığı betimsel tarama türünde uygulamalı bir araştırma niteliğindedir. Uygulamalı araştırmalar üretilmiş veya üretilmekte olan kuramsal bilginin denemeli uygulamasıdır ve var olan uygulamaları iyileştirme yönünde somut katkılar sunabilmektedir (Karasar, 2012).

Araştırmanın ikinci aşamasında simülasyonlar sonucunda belirlenen en uygun BOBUT koşulları yardımıyla ALES'in canlı BOBUT uygulaması yapılmıştır. ALES canlı BOBUT uygulamasının ALES kağıt-kalem testi ile ilişkisi incelendiği için çalışmanın ikinci aşaması ilişkisel araştırma niteliğindedir. İlişkisel araştırmalar aynı bireylerden elde edilen farklı değişkenlere ait ölçümlerin arasındaki ilişkiyi incelemek için kullanılan araştırmalardır (Mertens, 2010).

Araştırmanın Çalışma Grubu

Bu araştırmada veriler iki ayrı gruptan elde edilmiştir. İlk çalışma grubunda kullanılan ALES verileri, ÖSYM'den temin edilmiştir. ÖSYM ile resmi yollar ile görüşmeler yapılmıştır bu izne ilişkin belge EK-2'de yer almaktadır. ÖSYM tarafından paylaşılan veriler, ALES 2021 yılında yapılan üç döneme (Dönem 1, Dönem 2, Dönem

3)' e ait olup katılımcı bilgileri gizlenmiş ve rastgele seçilen 10.000 katılımcının sayısal ve sözel alt testlerde verdikleri yanıtlar 1 (doğru) ve 0 (yanlış) olarak kodlanmıştır. Bu gruptaki veriler üzerinde MTK varsayımları incelenmiş ve MTK'ya dayalı madde ve yetenek parametre kestirimleri gerçekleştirilmiştir. Ardından, elde edilen bulgular doğrultusunda bu grup üzerinde post-hoc simülasyon çalışmaları yürütülmüştür.

Çalışma Grubu 1

Tablo 3

2021 ALES Dönemlerine Göre Katılımcı Bilgileri

Dönem	Katılımcı Sayısı	Erkek Sayısı (%)	Kadın Sayısı (%)	Yaş Ortalaması	Standart Sapma
1	3500	1478 (%42.2)	2022 (%57.8)	27.6	5.9
2	3500	1591 (%45.5)	1909 (%54.5)	27.5	6.3
3	3000	1321 (%44.0)	1679 (%56.0)	27.4	6.2

Tablo 3'te, 2021 yılı ALES'in üç dönemine ait katılımcı sayıları, cinsiyet dağılımları ve yaş ortalamaları gösterilmektedir. Her üç dönemde de kadın katılımcı sayısı erkek katılımcı sayısından fazladır. Birinci dönemde kadın katılımcı oranı %57,8, ikinci dönemde %54,5 ve üçüncü dönemde %56,0 olarak hesaplanmıştır. Erkek katılımcı oranları ise sırasıyla %42,2, %45,5 ve %44,0'dır. Yaş ortalamalarına bakıldığında üç dönem arasında büyük farklılık bulunmamaktadır.

Çalışma Grubu 2

ALES-BOBUT uygulaması ile ALES'in kâğıt-kalem formundan elde edilen sonuçları karşılaştırmak amacıyla, 101 katılımcıdan oluşan bir çalışma grubuyla iki aşamalı bir veri toplama süreci yürütülmüştür. Çalışma grubunun oluşturulmasında, ALES'in farklı akademik alanlardan gelen bireylere yönelik bir sınav olması göz önünde bulundurulmuştur. Bu doğrultuda, katılımcıların alan çeşitliliğini yansıtacak şekilde farklı fakülte ve bölümlerden seçilmesine özen gösterilmiş ve çalışma grubunun alan bazında heterojen bir yapıya sahip olması amaçlanmıştır. Çalışma grubu, Adıyaman Üniversitesi bünyesindeki farklı fakülte, bölüm ve sınıflarda öğrenim gören öğrencilerden oluşmaktadır. Araştırma verileri 2025–2026 eğitim-öğretim yılında toplanmış olup,

uygulamaya katılan öğrencilerin öğrenim gördükleri bölüm ve sınıf düzeyine göre dağılımları Tablo 4'te sunulmuştur.

Tablo 4

ALES Canlı BOBUT ve ALES Kağıt-Kalem Uygulamasına Katılan Katılımcı Bilgileri

Bölüm/Sınıf Düzeyi	**Cinsiyet**			
	Kız	**Erkek**	**Toplam**	**Yüzde**
Psikolojik Danışmanlık ve Rehberlik/4	13	2	15	14.85
İktisat/2	10	2	12	11.88
İşletme/4	7	5	12	11.88
İlköğretim Matematik Öğretmenliği/3	13	9	22	21.78
İlköğretim Matematik Öğretmenliği/2	11	3	14	13.86
Fen Edebiyat Matematik/1-4	7	9	16	15.84
Bilgisayar Programcılığı/2	0	10	10	9.90
Toplam	**61**	**40**	**101**	**100**

Tablo 4'te yer alan iki uygulamaya katılan katılımcıların cinsiyet dağılımı incelendiğinde, toplam 61 kız ve 40 erkek öğrenci yer almaktadır. Alan yazında yapılan kağıt-kalem ve BOBUT uygulamalarının karşılaştırıldığı çalışmalarda, farklı araştırmacılar farklı örneklem büyüklükleri kullanmıştır ve örneklem büyüklüğünün kaç olması gerektiği konusunda ortak bir görüş bulunmamaktadır. Daha önce yapılan çalışmalar incelendiğinde, örneklem büyüklüklerinin 25 ile 242 arasında değiştiği görülmektedir (Alkan, 2021; Arslan vd., 2022; Aybek ve Çıkrıkçı, 2018; Bringsjord, 2001; Chew, 1997; de Beer ve Visser, 1998; English vd., 1997; Kalender ve Berberoğlu, 2017; Özbaşı ve Demirtaşlı, 2015; Şenel, 2017; Žitný vd., 2012). Bu durum, kağıt-kalem ve BOBUT uygulamalarında örneklem büyüklüğünün belirlenmesinin, araştırmanın uygulanabilirliği, bağlamı ve amaçlarına bağlı olarak değişebileceğini göstermektedir. ALES kağıt-kalem testinde yer alan maddelerin sayısının 100 olması ve testin kağıt-kalem formunun uygulanma süresinin 2.5 saat olması gibi pratik sebepler göz önünde bulundurularak, çalışmanın analizleri 101 katılımcıyla yürütülmüştür.

Verilerin Toplanması

Bu bölümde, araştırmada kullanılan veri toplama aracı olan ALES'e ve veri toplama sürecine ilişkin bilgilere yer verilmiştir.

Veri Toplama Aracı

Araştırmada veri toplama aracı olarak ÖSYM tarafından akademik personel ve lisansüstü eğitime giriş için hazırlanan ALES testi kullanılmıştır. Bu sınav ÖSYM tarafından yapılan merkezi bir sınavdır ve ilk kez 1997 yılında Lisansüstü Eğitimi Giriş Sınav (LES) adı altında uygulanmaya başlanmıştır. 2007 yılında ise Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı (ALES) olarak güncellenmiş ve hala aynı ad altında uygulanmaktadır (ÖSYM, 2023). Uygulandığı ilk günden beri sayısal ve sözel becerilerini ölçen alt testlerden oluşmaktadır. Bu alt testler 2007-2010 tarihleri arasında sayısal-1 40 madde sayısal-2 40 madde, sözel testi 80 madde olmak üzere üç alt testten, 2011-2013 tarihleri arasında sayısal-1 50 madde sayısal-2 50 madde, sözel-1 50 madde sözel-2 50 madde olmak üzere dört alt testten, 2014- 2017 tarihleri arasında sayısal-1 40 madde sayısal-2 40 madde, sözel-1 40 madde sözel-2 40 madde olmak üzere dört alt testten oluşmaktadır. Bu sınavlar 2017 yılına kadar yılda iki defa uygulanmıştır. 2018'den itibaren ise sözel testi ve sayısal testi olmak üzere iki alt testten oluşmakta ve yılda üç defa uygulanmaktadır. Sayısal alt testi, adayların sayısal ve mantıksal akıl yürütme (muhakeme) becerilerini ölçmeye yönelik 50 çoktan seçmeli maddeden; sözel alt testi, sözel akıl yürütme (muhakeme) becerilerini ölçmeye yönelik 50 çoktan seçmeli maddeden oluşmaktadır (ÖSYM, 2018a). 100 maddeden oluşan bu sınav için 150 dakika (2,5 saat) cevaplama süresi verilmektedir (ÖSYM, 2018b). Maddeler, bireylerin mezun oldukları bölümlerde kazanılan yeterlikleri ve bilgileri ölçmeye yönelik değil farklı alanlardan gelen bireylerin cevaplayabilecekleri niteliktedir (ÖSYM, 2018a; ÖSYM, 2018b). Bu maddeler 1-0 şeklinde ikili puanlanan maddelerdir. ALES sınavında sayısal, sözel ve eşit ağırlık olmak üzere üç puan türü hesaplanmaktadır. ÖSYM ALES sınavlarının gizliliğinden dolayı, 2022 yılından itibaren soru kitapçığında yer alan maddelerin %10'unu kamuoyuyla paylaşmaktadır. Bunun yanı sıra ÖSYM sınava giren adaylara sınav kitapçığında yer alan tüm maddelere on gün süreyle erişim izni vermektedir.

ÖSYM ALES'in uygulanması sonucunda elde ettiği verilere dayalı yayınladığı raporlarda testin çeşitli psikometrik özelliklerini yayınlamaktadır. Sınava giren aday sayısının cinsiyete göre dağılımı, alt testlerin ortalaması, ortalama ayırt edicilik gücü, ortalama güçlük düzeyi ve alt testlere ait güvenirlik katsayıları raporda yer almaktadır. ÖSYM'nin kamuoyuyla paylaştığı son rapor olan 2018 ALES/2 raporunda sayısal alt teste ait ortalama güçlük 0.38, ortalama ayırt edicilik 0.57, iç tutarlık güvenirlik katsayısı

0.95; sözel alt teste ait ortalama güçlük 0.58, ortalama ayırt edicilik 0.43, iç tutarlık güvenirlik katsayısı 0.91 olarak belirtilmiştir (ÖSYM, 2018b). Bu araştırma kapsamında 10.000 katılımcıdan elde edilen verilerle ALES'in sayısal ve sözel alt testlerine ait Cronbach Alfa güvenirlik katsayıları hesaplanmıştır. Elde edilen sonuçlara göre, ALES 1, ALES 2 ve ALES 3 dönemlerine ilişkin sayısal alt testlerin güvenirlik katsayıları sırasıyla 0.97, 0.97 ve 0.96; sözel alt testlerin güvenirlik katsayıları ise 0.84, 0.89 ve 0.86 olarak bulunmuştur. Tüm alt testlere ilişkin Cronbach Alfa değerlerinin 0.80'in üzerinde olduğu görülmektedir. Bu bulgular doğrultusunda, ALES alt testlerinden elde edilen puanların yüksek derecede güvenilir olduğu söylenebilir.

Verilerin Çözümlenmesi

Verilerin analizinde SimulCAT, R, SPSS ve Microsoft Excel programları kullanılarak araştırma amaçları doğrultusunda gerekli analizler gerçekleştirilmiştir.

Ön İncelemeler

Araştırmada ilk olarak ALES 2021 yılına ait üç form için kayıp değer ve uç değer gibi ön analizler SPSS programı yardımıyla düzenlenmiştir. ÖSYM'den temin edilen veri dosyasında, Y1–Y50 arasındaki maddeler sayısal alt teste, Y51–Y100 arasındaki maddeler ise sözel alt teste karşılık gelmektedir. Bu nedenle çalışmada tüm maddeler "Y" harfi ile başlayan bir biçimde kodlanmıştır.

Veri seti incelendiğinde 2021 ALES dönem 1 sayısal alt testinde madde bazında en yüksek %73,3 ve en düşük %14,1, sözel alt testinde madde bazında en yüksek %49,3 ve en düşük %3,1; 2021 ALES dönem 2 sayısal alt testinde madde bazında en yüksek %84,2 ve en düşük 13,1, sözel alt testinde madde bazında en yüksek %68,1 ve en düşük %4,3; dönem 3 sayısal alt testinde madde bazında en yüksek %74,8 ve en düşük %11,1, sözel alt testinde madde bazında en yüksek %47,8 ve en düşük %3,6 kayıp veri bulunmaktadır. Üç döneme ait detaylı kayıp veri sayıları ve oranları EK 3, EK 4 ve EK 5'te yer almaktadır.

Veri setinde kayıp veri bulunduğundan dolayı ilk olarak kayıp veri oranları sayısal ve sözel alt testlerde ayrı ayrı incelenmiştir. Kayıp veriler MTK modellerinin performansı göz önünde bulundurulacak şekilde çıkarılmıştır. 1PL modelin uygulanabilmesi için en az 200 kişilik bir örneklem büyüklüğünün yeterli olduğu belirtmektedir (Wright ve Stone,

1979). Bazı araştırmalar 2PL ve 3PL modelinin çok daha büyük örneklemlere (örneğin 1.000 birey) ihtiyaç duyulduğu (Reckase, 1979), a ve c parametresinin kestirimde 660 ve üzeri (Fotaris vd., 2010) katılımcının gerektiğini, 200-1000 örneklem büyüklüğünün kestirimlerde benzer sonuçlar verdiğini belirtmektedir (Linacre, 1994). BOBUT için bir madde havuzundaki maddeleri kalibre etmek amacıyla ise 500 üzeri örneklem büyüklüğü gerektiği (Reeve ve Fayers, 2005) belirtilmektedir. Bu nedenle, örneklem büyüklüğü 700 bireyin altına düşmeyecek şekilde kayıp veriler satır bazında silinmiştir. Kayıp veriler ayıklandıktan sonra uç değer incelemesi z puanları yardımıyla hesaplanmıştır. Dönem 1 sayısal alt testinde 11, sözel alt testinden 3; dönem 2 sayısal alt testinde 2, sözel alt testinden 3; dönem 3 sayısal alt testinde 9, sözel alt testinden 10 uç değer silinmiştir. Son durumda ALES Dönem 1 sayısal alt testte 715 birey, sözel alt testte 1481; ALES Dönem 2 sayısal alt testte 704, sözel alt testte 1244; ALES Dönem 3 sayısal alt testte 708, sözel alt testte 1598 birey bulunmaktadır.

Kayıp verilerin Tamamen Seçkisiz Kayıp (TSK-MCAR- Missing Completely at Random) olup olmadığını kontrol etmek için Little's MCAR Testi kullanılmıştır. Bir veri setinde eksik verilerin TSK olarak kabul edilebilmesi için, kayıp olma olasılığının ne değişkenin kendi değerleriyle ne de diğer değişkenlerle ilişkili olmaması gerekir (Little ve Rubin, 1987). Test sonucunda elde edilen kestirimler $p<.05$ olduğu durumlarda mevcut olan verinin TSK olmadığı seçkisiz kayıp (MAR-Missing at Random) veya seçkisiz olmayan kayıp (MNAR-Missing Not at Random) sonucuna varılır (Little, 1998). Dönem 1, 2 ve 3'e ait sayısal ve sözel alt testler için yapılan Little's MCAR testi sonuçları incelenmiştir. Elde edilen bulgular aşağıda Tablo 5'te verilmiştir:

Tablo 5

Dönemlere Göre Alt Testlerin Little's MCAR Testi Sonuçları

Dönem	Alt Test	χ^2 (df)	Anlamlılık
Dönem 1	Sayısal	$\chi^2(20734) = 20160.844$	$p > .05$
Dönem 1	Sözel	$\chi^2(19313) = 20039.552$	$p < .05$
Dönem 2	Sayısal	$\chi^2(20616) = 20940.286$	$p > .05$
Dönem 2	Sözel	$\chi^2(18728) = 19366.538$	$p < .05$
Dönem 3	Sayısal	$\chi^2(16196) = 16975.991$	$p < .05$
Dönem 3	Sözel	$\chi^2(22566) = 23692.954$	$p < .05$

Tablo 5 incelendiğinde, yalnızca Dönem 1 ve Dönem 2 Sayısal alt testlerinde p değerleri .05'in üzerinde çıkmış olup, bu alt testlerde eksik verilerin TSK olduğu söylenebilir. Buna karşılık Dönem 1 Sözel, Dönem 2 Sözel, Dönem 3 Sayısal ve Dönem

3 Sözel alt testlerinde p değerleri .05'ten küçük olduğundan, bu testlerde eksik verilerin TSK olmadığı sonucuna ulaşılmıştır. Bu doğrultuda, genel olarak değerlendirildiğinde eksik verilerin tamamının TSK varsayımını sağlamadığı, bu sebeple kayıp veri atamasına karar verilmiştir.

Kayıp veri temizliğinden sonra ALES dönem 2 sayısal alt testinde kayıp veri oranı %15'ten yüksek maddeler bulunduğu görülmüştür. Diğer dönem ve alt testlerde madde bazında %15 ve altında kayıp veriler mevcuttur. Bu durumda, kayıp veri ataması yapılmasına karar verilmiştir. Chen vd., (2020) %30 kayıp veri oranı ile Beklenti-Maksimizasyon (EM) yöntemini kullanarak eksik veriyi tamamlamanın mümkün olduğunu ve bu durumda yanlılık ile RMSE düzeylerinin kabul edilebilir sınırlar içinde kaldığını belirtirken, Smith vd., (2012) kayıp veri oranı %40'a kadar çıktığında bile EM ile atanan kayıp verinin güç hassasiyetini etkilemediğini belirtmiştir. Bu sebeple bu çalışmada, ALES verilerinde bulunan kayıp veriler için EM algoritması kullanılmıştır. Kayıp verilerle başa çıkmada EM algoritması, En Çok Olabilirlik (Maximum Likelihood, ML) yaklaşımı temelinde çalışan ve eksik verilerin dağılımını dikkate alarak kestirim yapan bir yöntemdir (Allison, 2009; Rubin, 1987; Schaffer, 1997). EM algoritması, Dempster vd., (1977) tarafından Seçkisiz kayıp veri (MAR- Missing at Random) varsayımı altında eksik verileri yönetmek amacıyla önerilmiştir. İki kategorili puanlanan maddelerden oluşan çoktan seçmeli testlerde, kayıp verilerin 'yanlış' olarak değerlendirilmesi (0 Atama yöntemi) ya da silinmesi gibi geleneksel yöntemlerin yanlı kestirimlere yol açtığından (DeMars, 2002; Finch, 2008; Linacre, 2004; Lord, 1983; Mislevy ve Wu, 1996) bu çalışmada kayıp verilerin en doğru şekilde işlenmesi amacıyla EM algoritması kullanılarak kayıp veri ataması yapılmasına karar verilmiştir. Özellikle, kayıp verilerin tam seçkisizlik özelliğini sağlamadığı ancak kısmî seçkisizlik gösterdiği durumlarda, EM algoritmasının ve En Çok Olabilirlik (ML) yöntemlerinin daha güvenilir ve titiz kestirimler ürettiği (Allison, 2002; Demir, 2013; Finch, 2008; Little ve Rubin, 1987; Schaffer, 1997) ve EM ile parametre tahminlerinin daha küçük standart hataları ile sonuçlandığı (Dong ve Peng, 2013; Springer ve Urban, 2014) literatürde sıklıkla vurgulanmaktadır. ALES gibi zaman sınırlaması olan testlerde bazı soruların boş bırakılması doğal bir durumdur. Özellikle testin sonlarına doğru, zaman yetersizliği, motivasyon düşüklüğü ya da bireylerin düşük yetenek düzeyine sahip olması gibi nedenlerle maddelerin boş bırakılması söz konusu olabilir. Bu tür durumlarda oluşan eksik veriler seçkisiz kayıp (MAR) veri özelliği taşıyabilir. Bu bağlamda, çalışmada

silmeye dayalı yöntemler veya 0 Atama yöntemi yerine, EM algoritması ile eksik veri ataması yapılmıştır.

Kağıt-kalem testlerinde ve bu testlerin bilgisayar ortamında sabit formlar hâlinde uygulandığı durumlarda, farklı test formlarından elde edilen puanların karşılaştırılabilirliğini sağlamak amacıyla test eşitleme yapılması gerekmektedir. BOBUT uygulamalarında ise, bireylerin farklı madde havuzları, farklı içerik dengelemesi ya da madde kullanım sıklığı koşullarında sınava girmesi durumunda elde edilen puanlar doğrudan karşılaştırılamayacağından eşitleme yapılması gerekirken; aynı madde havuzu ve aynı uygulama koşullarında gerçekleştirilen BOBUT uygulamalarında test eşitlemesine genellikle gerek duyulmamaktadır (Wang ve Kolen, 2001; Kolen ve Brennan, 2014). Bu çalışmada kullanılan BOBUT uygulamasında tek bir madde havuzu, sabit içerik dengelemesi ve madde kullanım sıklığı koşulları bulunmakta olup, her birey için maddeler kendi yetenek düzeyine uygun biçimde test sırasında aynı koşullar altında rastgele seçilmektedir. Dolayısıyla her bireyin yeteneği, test sürecinde sunulan maddelere dayalı olarak dinamik şekilde tahmin edilmekte; çalışmanın amacı farklı madde havuzları ya da ALES'in farklı formlarından elde edilen puanların birbirinin yerine kullanılmasını gerektirmediğinden, formlar arasında test eşitlemesi yapılmasına gerek duyulmamıştır

MTK Temel Varsayımlar ve Model Veri Uyumu İçin İncelemeler

Model-veri uyumu, MTK çerçevesinde modelin birey ve maddelere ilişkin verileri ne ölçüde doğru temsil ettiğini değerlendiren bir analizdir. Herhangi bir modelin verilere kesin uyumunu belirleyebilecek mutlak bir ölçüt bulunmamakla birlikte, model-veri uyumunu değerlendirmek için tek boyutluluk, yerel bağımsızlık gibi MTK'nın temel varsayımların test edilmesi gerekir (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1989). MTK'nın varsayımlarından olan tek boyutluluk için faktör analitik yöntemler, özdeğer ve yamaç birikinti grafiği (scree plot) incelemesi yaygın olarak kullanılmaktadır (Reckase,1979; Reise ve Revicki, 2015; Revicki vd., 2014). Bu sebeple ALES alt testlerinin MTK'nın tek boyutluluk varsayımını sağlayıp sağlamadığını incelemek amacıyla doğrulayıcı faktör analizi (DFA), özdeğer incelemesi ve paralel analiz yapılmıştır. DFA analizi, R programında *"lavaan"* (Rosseel, 2012) paketi yardımıyla yapılmıştır.

DFA için çeşitli uyum indeksleri bulunmaktadır (Cole, 1987). Bunlar içinde en sık kullanılanları Ki-Kare Uyum Testi ($\chi 2$), GFI (Goodness of Fit Index), AGFI (Adjusted

Goodness of Fit Index), RMSEA (Root Mean Square Error of Approximation), SRMR (Standardized Root Mean Square Residual), TLI (Tucker-Lewis Index) ve CFI (Comparative Fit Index) gibi uyum indeksleridir (Garrido vd., 2016; Hu ve Bentler, 1999). Ki-karenin 2 ile 3 arasındaki bir oranda olması, iyi bir veri-model uyumunu gösterir. (Bollen, 1989). RMSEA değerinin .05'in altında olmasını iyi uyumun göstergesi .05 ile .08 arasında olmasının orta düzeyde uyumu, .10'un üzerindeki değerlerin ise kötü uyumu ifade ettiği belirtilmektedir (Browne ve Cudeck, 1993). SRMR değerleri .08'in altında olduğunda, verinin modelle iyi bir uyumda olduğu söylenebilir (Hu ve Bentler, 1999). CFI ve TLI indeksleri 0 ile 1 arasında değişir, daha yüksek değerler modelin büyük uyum iyileşmelerini gösterir. CFI ve TLI değerleri .90 ve .95'in üzerinde olduğunda, veriye kabul edilebilir ve mükemmel uyumu yansıttığı kabul edilebilir (Hu ve Bentler, 1999; Yu, 2002). Son olarak GFI ve AGFI değerlerinin 0.90'den yüksek olması modelin verilerle uyumuna yönelik birer kriter olarak da değerlendirilmekle (Anderson ve Gerbing, 1984; Cole, 1987) birlikte bu değerler için genel bir kural, .90 ve üzerindeki değerlerin iyi uyumu, .85'in üzerindeki değerlerin ise kabul edilebilir uyumu gösterdiğidir (Schermelleh-Engel vd., 2003). ALES sayısal ve sözel alt testlere ilişkin yapılan DFA uyum sonuçları Tablo 6' da verilmiştir.

Tablo 6

Alt Testlere İlişkin DFA Uyum İndeksleri

Alt Test	$\chi 2$/sd	RMSEA [95% GA]	SRMR	CFI	TLI	GFI	AGFI
Dönem 1 Sayısal	2.37	0.044 [0.042–0.046]	0.041	0.848	0.842	0.859	0.847
Dönem 2 Sayısal	2.23	0.042 [0.039–0.044]	0.040	0.863	0.857	0.869	0.858
Dönem 3 Sayısal	2.76	0.050 [0.048–0.052]	0.042	0.841	0.834	0.835	0.821
Dönem 1 Sözel	1.96	0.024 [0.025–0.027]	0.032	0.844	0.837	0.939	0.934
Dönem 2 Sözel	2.34	0.033 [0.031–0.034]	0.040	0.820	0.812	0.910	0.902
Dönem 3 Sözel	2.27	0.028 [0.027–0.030]	0.034	0.823	0.815	0.930	0.924

GA: Güven Aralığı

Tablo 6'da verilen sayısal ve sözel alt testlere ait $\chi 2$/sd değerleri incelendiğinde 1.96 ile 2.76 arasında değiştiği görülmektedir. Bu durum tüm dönemlerde modelin veriyle kabul edilebilir düzeyde uyum sağladığını göstermektedir. Sayısal ve sözel alt testlerde

RMSEA ve SRMR değerlerinin tüm dönemlerde .05'in altında olması iyi uyuma işaret ederken, CFI ve TLI değerlerinin .90'ın altında kalması model uyumunun sınırlı olduğunu göstermektedir. Ancak RMSEA'nın iyi düzeyde olması, modelin kabul edilebilir olduğunu desteklemektedir. Çünkü CFI ve TLI uyum ölçütleri belirli bir derecede düşük ancak RMSEA değeri daha iyi bir uyum gösteriyorsa, bu durum modelin kabul edilebilir bir uyum sağladığını ve daha fazla endişe gerektirmediğini gösterebilir. Ancak, her iki uyum ölçütünün de zayıf olduğu bir durumda, bu, modelin genel uyumunun yetersiz olduğunu ve modelin doğruluğunun ciddi şekilde sorgulanması gerektiğini işaret eder (Kenny ve McCoach, 2003). GFI ve AGFI değerlerinin genelde .85'in üzerinde olması, modelin kabul edilebilir bir uyum sağladığını desteklemektedir. Bu sonuçlar hem sayısal hem de sözel alt testlerin tek boyutluluk varsayımını genel olarak sağladığını göstermektedir. ALES'in sayısal ve sözel alt testlerinin genel olarak MTK'nın varsayımı olan tek boyutluluğu kabul edilebilir düzeyde karşıladığı görülmüştür. Model veri uyumundan sonra maddelerin bulunduğu alt testle olan ilişkisini de değerlendirmek için madde faktör yükleri ve manidarlık (p) düzeyleri incelenmiştir. Sayısal alt testlerde yer alan maddelere ait faktör yükleri ve manidarlık düzeyleri Tablo 7'de gösterilmiştir.

Tablo 7

Sayısal Alt Testlere Ait DFA Faktör Yükleri ve Manidarlık Düzeyleri

Madde	Dönem 1		Dönem 2		Dönem 3	
	Faktör Yükü	p	Faktör Yükü	p	Faktör Yükü	p
Y1	0.444		0.554		0.572	
Y2	0.550	0.000	0.658	0.000	0.670	0.000
Y3	0.610	0.000	0.701	0.000	0.662	0.000
Y4	0.680	0.000	0.665	0.000	0.522	0.000
Y5	0.195	0.000	0.710	0.000	0.445	0.000
Y6	0.618	0.000	0.583	0.000	0.260	0.000
Y7	0.500	0.000	0.673	0.000	0.691	0.000
Y8	0.421	0.000	0.644	0.000	0.331	0.000
Y9	0.434	0.000	0.480	0.000	0.596	0.000
Y10	0.580	0.000	0.628	0.000	0.611	0.000
Y11	0.669	0.000	0.510	0.000	0.708	0.000
Y12	0.660	0.000	0.546	0.000	0.686	0.000
Y13	0.617	0.000	0.725	0.000	0.549	0.000

(Devam ediyor)

Tablo 7 (Devam)

Sayısal Alt Testlere Ait DFA Faktör Yükleri ve Manidarlık Düzeyleri

Madde	Dönem 1		Dönem 2		Dönem 3	
	Faktör Yükü	p	Faktör Yükü	p	Faktör Yükü	p
Y14	0.659	0.000	0.565	0.000	0.398	0.000
Y15	0.611	0.000	0.638	0.000	0.580	0.000
Y16	0.674	0.000	0.634	0.000	0.592	0.000
Y17	0.537	0.000	0.539	0.000	0.483	0.000
Y18	0.420	0.000	0.650	0.000	0.642	0.000
Y19	0.482	0.000	0.621	0.000	0.503	0.000
Y20	0.479	0.000	0.368	0.000	0.750	0.000
Y21	0.494	0.000	0.484	0.000	0.360	0.000
Y22	0.457	0.000	0.577	0.000	0.347	0.000
Y23	0.370	0.000	0.580	0.000	0.569	0.000
Y24	0.525	0.000	0.653	0.000	0.636	0.000
Y25	0.382	0.000	0.465	0.000	0.279	0.000
Y26	0.170	0.000	0.451	0.000	0.515	0.000
Y27	0.471	0.000	0.616	0.000	0.602	0.000
Y28	0.418	0.000	0.579	0.000	0.510	0.000
Y29	0.528	0.000	0.680	0.000	0.606	0.000
Y30	0.547	0.000	0.386	0.000	0.608	0.000
Y31	0.673	0.000	0.663	0.000	0.477	0.000
Y32	0.684	0.000	0.553	0.000	0.614	0.000
Y33	0.419	0.000	0.540	0.000	0.703	0.000
Y34	0.603	0.000	0.239	0.000	0.568	0.000
Y35	0.549	0.000	0.609	0.000	0.499	0.000
Y36	0.495	0.000	0.596	0.000	0.683	0.000
Y37	0.599	0.000	0.562	0.000	0.587	0.000
Y38	0.595	0.000	0.656	0.000	0.273	0.000
Y39	0.492	0.000	0.708	0.000	0.361	0.000
Y40	0.295	0.000	0.253	0.000	0.459	0.000
Y41	0.628	0.000	0.619	0.000	0.490	0.000
Y42	0.577	0.000	0.506	0.000	0.491	0.000
Y43	0.528	0.000	0.756	0.000	0.450	0.000
Y44	0.584	0.000	0.707	0.000	0.626	0.000
Y45	0.650	0.000	0.688	0.000	0.601	0.000
Y46	0.435	0.000	0.593	0.000	0.557	0.000
Y47	0.607	0.000	0.699	0.000	0.537	0.000
Y48	0.540	0.000	0.526	0.000	0.573	0.000
Y49	0.353	0.000	0.443	0.000	0.460	0.000
Y50	0.419	0.000	0.417	0.000	0.575	0.000

Sayısal alt testlere ait DFA sonuçlarına göre, üç dönemde de incelenen tüm maddeler istatistiksel olarak anlamlıdır ($p < 0.05$). Bu durum, maddelerin ait oldukları faktörle anlamlı bir ilişki içinde olduğunu göstermektedir. Faktör yükleri açısından değerlendirildiğinde, genellikle .30'ın üzerinde değerler elde edilmiştir, bu da madde ve

faktör ilişkilerinin yeterli düzeyde olduğunu göstermektedir. Ancak bazı maddelerin (Dönem 1: Y5, Y26; Dönem 2: Y34, Y40) faktör yükleri .30'dan düşük ve bazı maddelerin (Dönem 1: Y40; Dönem 3: Y6, Y25, Y38) .30 civarında yüke sahip olduğu sonucuna ulaşılmıştır. Bu nedenle, söz konusu maddelerin parametreleri ayrıntılı olarak incelenmiştir. Faktör yükü düşük olan ve aynı zamanda ayırt edicilik düzeyi çok düşük bulunan maddeler testten çıkarılmıştır. Ayırt edicilik düzeyinin çok düşük olduğu maddeler, Baker (2001) tarafından belirtilen kriterlere göre 0.35 ve 0.35'in altında olan maddeler olarak tanımlanmıştır. Madde ayırt edicilik değeri 1'den büyükse, maddenin yüksek ayırt ediciliğe sahip olduğu söylenebilir (Embretson ve Reise, 2000; deMars, 2010). Baker (2001) ise madde ayırt ediciliği değerlerine yönelik farklı bir sınıflandırma yapmıştır. Buna göre, madde ayırt edicilik değeri 0.01 ile 0.34 arasındaki maddeler "çok düşük", 0.35 ile 0.64 arasındaki maddeler "düşük", 0.65 ile 1.34 arasındaki maddeler "orta", ve 1.35 ve üzerindeki maddeler ise "yüksek" ayırt edicilik göstermektedir. Şans parametresi olan c-parametresinin ise $0 \leq c \leq 0.35$ aralığında olmasını önermektedir. Madde parametre kestirim sonuçları, Baker (2001) tarafından belirtilen ayırt edicilik ve şans parametre değerleri referans alınarak yorumlanmıştır. Faktör yükü düşük olmasına rağmen ayırt ediciliği .35 üstü olan maddelerin faktörle ilişkisinin manidarlığı da göz önünde bulundurulduğunda testte yer almasına karar verilmiştir. Buna göre faktör yükleri düşük olan dönem 1 sayısal alt testine ait Y5 (a=0.918) ve Y26 (a=0.831) maddelerinin ayırt edicilikleri yeterli düzeyde bulunduğu için testte yer alması uygun görülmüştür. Tüm dönemlerde bazı maddelerin yükleri dalgalansa da genel olarak modelin iç tutarlılığı yüksek olup, DFA kapsamında maddelerin büyük kısmı güçlü ve anlamlı faktör yüklerine sahiptir. Sözel alt testlerde yer alan maddelere ait faktör yükleri ve manidarlık düzeyleri Tablo 8'de gösterilmiştir.

Tablo 8

Sözel Alt Testlere Ait DFA Faktör Yükleri ve Manidarlık Düzeyleri

Madde	Dönem 1		Dönem 2		Dönem 3	
	Faktör Yükü	p	Faktör Yükü	p	Faktör Yükü	p
Y51	0.119		0.256		0.187	
Y52	0.117	0.003	0.457	0.000	0.167	0.000
Y53	0.066	0.040	0.318	0.000	0.214	0.000
Y54	0.134	0.002	0.240	0.000	0.250	0.000
Y55	0.291	0.000	0.280	0.000	0.084	0.004
Y56	0.250	0.000	0.367	0.000	0.290	0.000
Y57	0.290	0.000	0.303	0.000	0.307	0.000

(Devam ediyor)

Tablo 8 (Devam)

Sözel Alt Testlere Ait DFA Faktör Yükleri ve Manidarlık Düzeyleri

Madde	Dönem 1		Dönem 2		Dönem 3	
	Faktör Yükü	p	Faktör Yükü	p	Faktör Yükü	p
Y58	0.372	0.000	0.397	0.000	0.288	0.000
Y59	0.383	0.000	0.365	0.000	0.154	0.000
Y60	0.322	0.000	0.339	0.000	0.307	0.000
Y61	0.437	0.000	0.426	0.000	0.275	0.000
Y62	-0.108	0.004	0.315	0.000	0.376	0.000
Y63	0.170	0.000	0.391	0.000	0.264	0.000
Y64	0.267	0.000	0.385	0.000	0.323	0.000
Y65	0.162	0.000	0.446	0.000	0.178	0.000
Y66	0.369	0.000	0.284	0.000	0.397	0.000
Y67	0.322	0.000	0.245	0.000	0.259	0.000
Y68	0.174	0.000	0.402	0.000	0.295	0.000
Y69	0.177	0.000	0.367	0.000	0.267	0.000
Y70	0.247	0.000	0.297	0.000	0.280	0.000
Y71	0.357	0.000	0.319	0.000	0.370	0.000
Y72	0.243	0.000	0.461	0.000	0.360	0.000
Y73	0.505	0.000	0.513	0.000	0.115	0.000
Y74	0.353	0.000	0.259	0.000	0.277	0.000
Y75	0.279	0.000	-0.096	0.000	0.292	0.000
Y76	0.265	0.000	0.383	0.000	0.219	0.000
Y77	0.327	0.000	0.379	0.000	0.253	0.000
Y78	0.367	0.000	0.355	0.000	0.296	0.000
Y79	0.331	0.000	0.331	0.000	0.290	0.000
Y80	0.427	0.000	0.396	0.000	0.280	0.000
Y81	0.170	0.000	0.469	0.000	0.043	0.116
Y82	0.054	0.082	0.277	0.000	0.334	0.000
Y83	0.426	0.000	0.542	0.000	0.333	0.000
Y84	0.399	0.000	0.260	0.000	0.346	0.000
Y85	0.364	0.000	0.301	0.000	0.370	0.000
Y86	0.322	0.000	0.439	0.000	0.252	0.000
Y87	0.071	0.030	0.375	0.000	0.213	0.000
Y88	0.398	0.000	0.326	0.000	0.380	0.000
Y89	0.334	0.000	0.418	0.000	0.472	0.000
Y90	0.190	0.000	0.467	0.000	0.482	0.000
Y91	0.304	0.000	0.416	0.000	0.441	0.000
Y92	0.141	0.000	0.449	0.000	0.399	0.000
Y93	0.438	0.000	0.421	0.000	0.472	0.000
Y94	0.386	0.000	0.364	0.000	0.582	0.000
Y95	0.386	0.000	0.403	0.000	0.544	0.000
Y96	0.455	0.000	0.385	0.000	0.484	0.000
Y97	0.407	0.000	0.399	0.000	0.399	0.000
Y98	0.467	0.000	0.337	0.000	0.416	0.000
Y99	0.368	0.000	0.388	0.000	0.375	0.000
Y100	0.510	0.000	0.411	0.000	0.435	0.000

Sözel alt testlere ait DFA sonuçlarına göre, üç dönemde de incelenen maddelerden 148 tanesi istatistiksel olarak anlamlıdır ($p < 0.05$). Dönem 1 sözel alt testinde yer alan Y82 (faktör yükü: 0.054) ve dönem 3 alt testinde yer alan Y81 (faktör yükü: 0.043) maddelerinin hem faktör yükü çok düşük olduğundan hem de istatistiksel olarak anlamlı olmadığından testten çıkarılmasına karar verilmiştir. Bu durum, ilgili maddelerin ait oldukları faktörle anlamlı bir ilişki içinde olmadığını göstermektedir. Faktör yükleri açısından değerlendirildiğinde, genellikle .30'ın üzerinde değerler elde edilmiştir, bu da madde ve faktör ilişkilerinin yeterli düzeyde olduğunu göstermektedir. Ancak bazı maddelerin faktör yükleri düşük (Dönem 1: Y90, Y92; Dönem 3: Y51, Y52 gibi) ve bazı maddelerin (Dönem 1: Y75; Dönem 2: Y70 gibi) .30 civarında yüke sahip olduğu sonucuna ulaşılmıştır. Bu nedenle, faktör yükü düşük maddelerin parametreleri ayrıntılı olarak incelenmiştir. Faktör yükü düşük olan ve aynı zamanda ayırt edicilik düzeyi de çok düşük bulunan maddeler testten çıkarılmıştır. Ancak, faktör yükü düşük olmasına rağmen ayırt ediciliği .35 üstü olan maddelerin faktörle ilişkisinin manidarlığı da göz önünde bulundurulduğunda testte yer almasına karar verilmiştir. Buna göre dönem 1 sözel alt testte yer alan Y62 (a=-0.284), Y68 (a=0.350), Y81 (a=0.340), Y87 (a=0.148), ve Y92 (a=0.262) maddelerinin ayırt edicilikleri çok düşük bulunduğu için testten çıkarılmıştır. Dönem 2 alt testinde yer alan Y75 (a=--0.190) ve dönem 3 alt testinde yer alan Y55 (a=0.316) maddelerinin ayırt edicilikleri çok düşük bulunduğu için testten çıkarılmıştır. MTK'nın tek boyutluluk varsayımının incelenmesi sonucunda, testin baskın bir tek faktörlü yapıyı sağladığı belirlenmiş ve bu bulguya dayalı olarak tek boyutlu MTK modeli altında parametre kestirimleri gerçekleştirilmiştir. Analiz sonuçları, sözel alt testinin tek boyutlu olduğunu desteklemekle birlikte, paragraf, sözel mantık ve cümle tamamlama gibi farklı maddelerin doğası gereği ikincil bilişsel süreçleri farklı düzeylerde harekete geçirebildiğini düşündürmektedir. Bu durumun, tek boyutlu MTK çerçevesinde bazı maddelerin ayırt edicilik parametrelerinin görece daha düşük kestirilmesine yol açmış olabileceği değerlendirilmektedir; testten çıkarılan maddelerin ölçme gücüne ilişkin bir zayıflık olarak değil, model varsayımları ve adaptif test bağlamı içinde ele alınması uygun görülmüştür.

Tek boyutluluk varsayımını incelemek için özdeğer incelemesi de yapılmıştır. Lord (1980), tek boyutluluk varsayımının değerlendirilmesinde birinci özdeğerin ikinci özdeğere oranının incelenmesini önermiştir. Özdeğer incelemesi R'da bulunan "*psych"* (Revelle, 2022) paketi yardımıyla yapılmış olup bu sonuçlar Tablo 9'da verilmiştir.

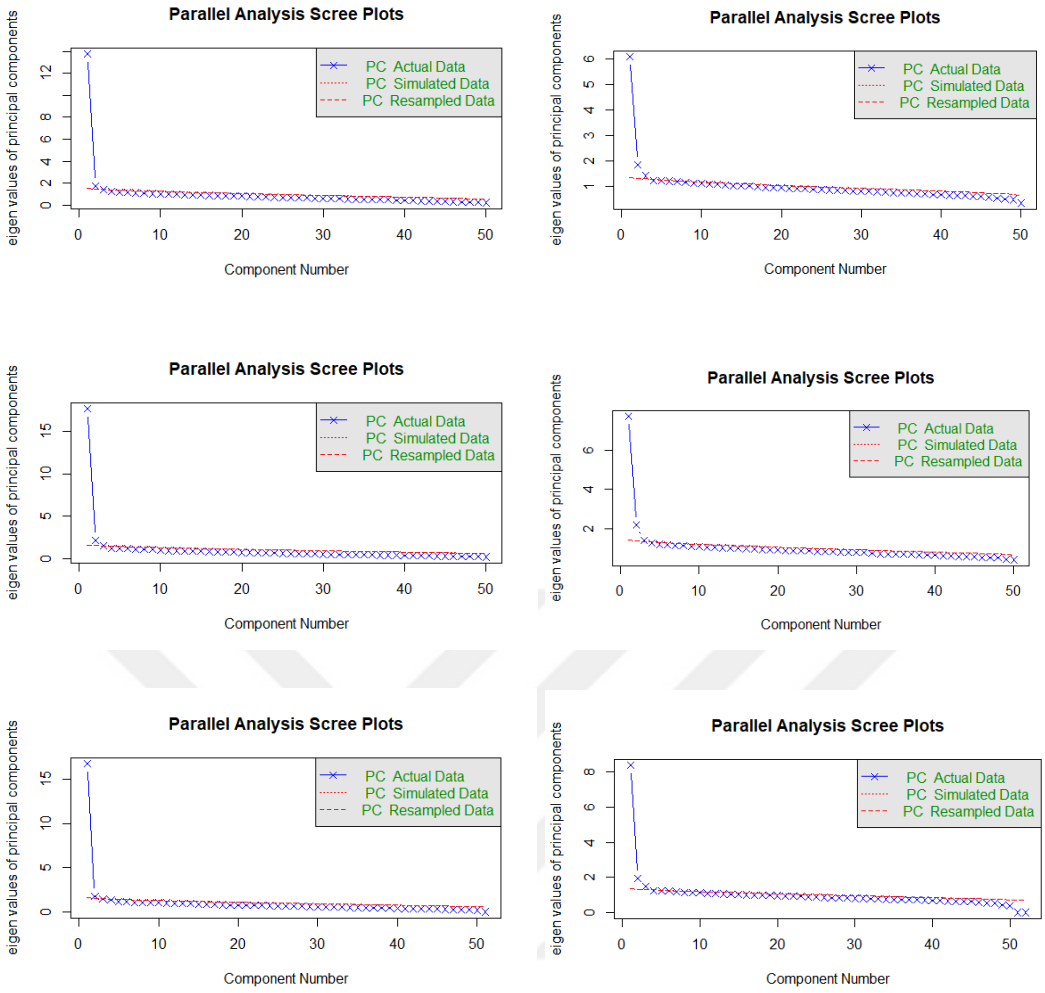
Tablo 9

ALES Alt Testleri Faktör Özdeğer Oranları

Alt test	λ1	λ2	λ3	$\frac{\lambda 1}{\lambda 2}$
Dönem 1 sayısal	13.73	1.77	1.45	7.75
Dönem 1 sözel	6.08	1.85	1.43	3.28
Dönem 2 sayısal	17.70	2.08	1.47	8.50
Dönem 2 sözel	7.70	2.20	1.38	3.50
Dönem 3 sayısal	16.72	1.75	1.42	9.55
Dönem 3 sözel	6.42	1.94	1.48	4.31

ALES alt testleri için hesaplanan özdeğerler incelendiğinde birinci özdeğerin 17.70 ile 6.08; ikinci özdeğerin 2.20 ile 1.75; üçüncü özdeğerin ise 1.38 ile 1.48 arasında değiştiği görülmektedir. Eğer birinci özdeğer çok baskınsa (çok büyükse) ve diğerleri birbirine yakın ve küçükse, ilk faktörün baskın olduğunu ve tek boyutluluk varsayımının desteklendiği ifade edilmiştir (Hattie, 1985). Elde edilen özdeğer oranları incelendiğinde birinci özdeğerlerin ikinci özdeğerlerden büyük olduğu, ikinci ve üçüncü özdeğerlerin birbirine yakın olduğu ve birinci özdeğerin ikinci özdeğere oranı incelendiğinde 3.28 ile 9.55 arasında değiştiği görülmektedir. Bu sonuçlar tek boyutluluk varsayımını olumlu yönde desteklemektedir.

Tek boyutluluk varsayımını değerlendirmek amacıyla ayrıca Horn (1965) tarafından geliştirilen Paralel Analiz tekniğinden yararlanılmıştır. Paralel Analiz boyutluluk belirleme konusunda en sık kullanılan ve önerilen (Kılıç ve Uysal, 2019, 2021; Lance vd., 2006; Worthington ve Whittaker, 2006) bir yöntemdir. Bu yöntem, gözlenen veri setinden elde edilen özdeğerlerin, aynı yapıya sahip tesadüfi (rastgele) veri setlerinden elde edilen özdeğerlerle karşılaştırılmasına dayanır ve ölçekte baskın bir faktörün var olup olmadığı incelenebilir. ALES alt testleri için paralel analiz R programında *psych* paketi yardımıyla yapılmıştır. Paralel analiz grafikleri Şekil 6'da verilmiştir.



Şekil 6. ALES Alt Testlerine Ait Paralel Analiz Grafikleri

Şekil 6 incelendiğinde ALES alt testleri için yapılan paralel analizler sonucunda tüm alt testler için başat faktörün bulunduğu görülmektedir. Paralel analiz sonuçlarında açıkça tek bir keskin düşüş görülmektedir. Bu durum, ALES'in her iki alt testte yer alan maddelerin tek ve başat bir faktörün altında gruplandığını işaret etmektedir.

MTK'nın diğer bir varsayımı olan yerel bağımsızlık varsayımı için sıklıkla kullanılan Yen (1984) Q3 istatistiği her alt test için ayrı ayrı incelenmiştir. Maddelerin tüm ikili kombinasyonlarını içeren bu istatistikte .20 üzeri korelasyonlar, yerel bağımsızlık varsayımının karşılanmadığını gösterir (DeMars, 2010). Tek boyutluluk varsayımının karşılanması durumunda yerel bağımsızlık varsayımının da karşılandığı alan yazında belirtilse de (de Ayala, 2009; Embretson ve Reise, 2000; Hambleton vd., 1991; Reise ve Revicki, 2015), yerel bağımsızlık varsayımı incelenmiştir. Q3 istatistikleri, R ortamında bulunan *"subscore"* paketi yardımıyla hesaplanmıştır. ALES

alt testlerine yönelik yapılan Yen'in Q3 istatistiği hesaplamasında, 291 madde ile 42.195 madde çiftine ilişkin Q3 istatistiklerinin büyük çoğunluğunun .20'den küçük, sayısal ve sözel alt testlerinde toplam 14 madde çiftinde ise 0.20 yakınlarında olduğu belirlendiğinden, yerel bağımsızlık varsayımının karşılandığı kabul edilmiştir.

Parametre kestirimlerinin yapılabilmesi için model-veri uyumunun değerlendirilmesi amacıyla, "*mirt"* paketi kullanılarak -2 Log Likelihood (-2LL), Akaike Bilgi Kriteri (AIC – Akaike Information Criterion), Bayesyen Bilgi Kriteri (BIC – Bayesian Information Criterion) ve modellerin uyumlarının karşılaştırmak için yapılan ANOVA sonuçlarının p değerleri hesaplanmıştır. Bu yaklaşım, modele ilişkin verilerin ne derece uyumlu olduğunu belirlemek ve parametrelerin güvenilir bir şekilde tahmin edilip edilmediğini değerlendirmek için tercih edilmiştir. AIC ve BIC değerleri model seçim konusunda parametre kestirimi yapacak en iyi modelin belirlenmesi alanında en çok bilinen yöntemlerdir (Chakrabarti ve Ghosh, 2011). Model karşılaştırmalarında, genellikle daha düşük -2LL, AIC ve BIC değerleri ile daha iyi model uyumu sağladığı düşünülür (Hu vd., 2016). ALES alt testlerine ilişkin model veri uyumları Tablo 10'da gösterilmiştir.

Tablo 10

ALES Alt Testlerine Ait Model Veri Uyumu Sonuçları

Alt test	Model	-2LL	AIC	BIC	p
Dönem 1 sayısal	2PL	-12148.89	24497.77	24955.00	p>.05
	3PL	-12132.69	24565.38	25251.22	
Dönem 2 sayısal	2PL	-15169.96	30539.91	30995.59	p<.05
	3PL	-15040.65	30381.30	31064.82	
Dönem 3 sayısal	2PL	-10618.74	21437.48	21893.72	p>.05
	3PL	-10590.95	21481.91	22166.28	
Dönem 1 sözel	2PL	-39415.82	79031.63	79561.68	p<.05
	3PL	-39340.75	78981.51	79776.58	
Dönem 2 sözel	2PL	-33178.25	66556.51	67069.12	p<.05
	3PL	-33041.68	66383.35	67152.26	
Dönem 3 sözel	2PL	-34872.02	69944.04	70481.69	p<.05
	3PL	-34825.38	69950.75	70757.23	

Tablo 10'da verilen model uyum değerleri incelendiğinde, -2LL, AIC ve BIC değerlerinin birçok alt testte 3PL lehine daha düşük çıktığı, ayrıca 2PL ve 3PL modelleri arasındaki karşılaştırmalarda bazı alt testlerde elde edilen anlamlı p-değerlerinin ($p < 0.05$) 3PL modelinin veriye daha iyi uyum sağladığını gösterdiği görülmektedir. Dönem

1 Sayısal ve Dönem 3 Sayısal testlerinde 2PL modeli daha uygun görünse de genel olarak alt testlerin 3PL modeline daha iyi uyum sağladığı söylenebilir. Kuramsal açıdan incelendiğinde ise, 3PL modelinin içerdiği tahmin parametresi (c) sayesinde, düşük yetenek düzeyine sahip bireylerin doğru cevaba şansla ulaşabilme olasılığı modellenebilmekte ve böylece daha gerçekçi madde karakteristik eğrileri elde edilebilmektedir (Baker, 2001; deMars, 2010; von Davier, 2009). Bu çalışmada kullanılan testlerin çoktan seçmeli yapısı ve düşük yetenek düzeyindeki bireylerin doğru cevap verme olasılıklarının sadece yetenek düzeyleriyle değil, aynı zamanda tahmin davranışlarıyla da ilişkili olabileceği göz önünde bulundurulduğunda, 3PL modelinin tercih edilmesi hem kuramsal hem de istatistiksel açıdan incelenerek tercih edilmiştir.

3PL modelinin uygunluğunun değerlendirilmesi amacıyla ayrıca maddelerin modelle olan uyum değerleri incelenmiştir. Bu amaçla maddeler için S-χ2 değerleri hesaplanmıştır. S-χ2, Orlando ve Thissen (2000) tarafından ikili puanlanan maddeler için geliştirilen madde-uyum indeksidir. Bu indeks ile toplam puandan elde edilen gözlenen ve beklenen değerler, χ2 istatistiği kullanılarak karşılaştırılır. Maddelerin model ile uyum gösterip göstermediği, 0.05 anlamlılık düzeyi kullanılarak belirlenir. Eğer uyum indeksinin p değeri 0.05'ten küçük ise madde modele uyumsuzdur. ALES'in her bir formunda yer alan maddeler için "*mirt* "paketi yardımıyla kestirilen S-χ2 değerleri sayısal alt testler için Tablo 11'de, sözel alt *testler için* ise Tablo 12'de verilmiştir.

Tablo 11

ALES Sayısal Alt Testlere Ait Madde-Model Uyum Değerleri

	Dönem 1		Dönem 2		Dönem 3	
Madde	S-χ2	p	S-χ2	p	S-χ2	p
Y1	35.78	0.12	31.05	0.15	5.81	0.21
Y2	18.28	0.69	27.70	0.68	16.31	0.70
Y3	20.11	0.83	44.03	0.14	8.67	0.65
Y4	14.91	0.31	36.07	0.24	20.10	0.69
Y5	13.73	0.91	35.75	0.43	17.48	0.83
Y6	28.30	0.45	46.35	0.14	32.09	0.23
Y7	12.86	0.91	51.73	**0.01**	21.24	**0.03**
Y8	32.85	0.43	47.47	0.12	27.74	0.42
Y9	22.20	0.73	47.44	0.20	17.07	0.81
Y10	36.12	0.09	36.95	0.52	31.29	0.09
Y11	11.69	0.82	29.61	0.59	15.89	0.26
Y12	27.06	0.46	32.94	0.74	11.32	0.73

(Devam ediyor)

Tablo 12 (Devam)

ALES Sayısal Alt Testlere Ait Madde-Model Uyum Değerleri

	Dönem 1		Dönem 2		Dönem 3	
Madde	S-χ2	p	S-χ2	p	S-χ2	p
Y13	23.54	0.60	41.65	0.05	15.79	0.83
Y14	17.54	0.94	25.96	0.96	31.70	0.14
Y15	26.86	0.26	17.93	0.98	36.68	**0.04**
Y16	21.82	0.70	35.64	0.62	21.22	0.57
Y17	20.76	0.60	34.35	0.19	19.13	0.75
Y18	31.59	0.17	43.41	0.19	26.82	0.26
Y19	30.09	0.22	55.26	**0.04**	19.06	0.75
Y20	19.71	0.76	60.36	**0.02**	5.38	0.61
Y21	28.49	0.60	29.99	0.82	18.03	0.52
Y22	21.80	0.75	38.51	0.49	21.42	0.37
Y23	34.51	0.12	37.08	0.56	21.47	0.26
Y24	30.45	0.17	26.71	0.87	24.87	0.36
Y25	26.21	0.56	39.56	0.49	12.58	0.96
Y26	25.62	0.54	39.55	0.49	29.51	0.20
Y27	43.65	**0.03**	43.76	0.24	20.09	0.39
Y28	15.54	0.93	39.87	0.39	24.01	0.52
Y29	16.42	0.95	45.68	0.13	29.94	**0.03**
Y30	38.78	0.09	29.16	0.87	18.20	0.64
Y31	22.51	0.37	26.95	0.52	32.29	0.12
Y32	27.79	0.37	45.17	0.27	21.41	0.37
Y33	18.65	0.81	32.07	0.78	16.02	0.72
Y34	33.56	0.07	43.04	0.38	23.19	0.51
Y35	39.28	0.06	36.16	0.28	27.09	0.35
Y36	17.92	0.88	31.69	0.67	14.23	0.82
Y37	28.19	0.35	47.72	0.19	34.77	**0.02**
Y38	23.36	0.61	33.36	0.45	29.81	0.16
Y39	30.12	0.36	31.87	0.42	26.05	0.52
Y40	19.99	0.87	47.63	0.09	33.17	0.13
Y41	37.90	**0.01**	47.65	0.14	21.43	0.72
Y42	21.90	0.69	29.94	0.88	18.29	0.83
Y43	17.57	0.94	25.39	0.71	27.30	0.16
Y44	20.99	0.74	30.47	0.49	21.32	0.56
Y45	22.40	0.67	20.62	0.95	11.92	0.94
Y46	15.45	0.97	34.53	0.54	16.13	0.71
Y47	36.86	0.05	30.22	0.74	12.30	0.91
Y48	19.56	0.85	32.50	0.80	15.76	0.78
Y49	21.72	0.89	42.54	0.45	20.66	0.71
Y50	14.03	0.99	40.86	0.48	11.09	0.99

Madde-model uyumu sağlamayan maddeler bold olarak gösterilmiştir.

ALES'in sayısal alt testlerinde yer alan 150 maddenin 3PL ile madde-model uyumunun incelenmesi amacıyla hesaplanan S-χ2 değerleri incelendiğinde 141 maddenin 3PL ile uyumlu olduğu (p>0.05), 9 maddenin ise madde-model uyumunu sağlamadığı (p<0.05) görülmüştür. ALES sayısal alt testlerinde yer alan maddelerin %94'ünün 3PL ile uyumlu olduğu sonucuna ulaşılmıştır.

Tablo 13

ALES Sözel Alt Testlere Ait Madde-Model Uyum Değerleri

	Dönem 1		Dönem 2		Dönem 3	
Madde	S-χ2	p	S-χ2	p	S-χ2	p
Y51	44.19	0.06	45.22	0.17	37.29	0.17
Y52	33.65	0.30	19.92	0.94	39.59	0.14
Y53	32.65	0.34	32.15	0.46	16.90	0.96
Y54	15.14	0.99	37.62	0.31	29.49	0.34
Y55	25.15	0.72	53.88	**0.03**	*	
Y56	24.77	0.78	39.49	0.28	19.11	0.94
Y57	16.47	0.98	27.99	0.72	27.80	0.37
Y58	44.40	**0.03**	41.35	0.21	25.91	0.68
Y59	38.35	0.12	31.38	0.64	38.37	0.14
Y60	27.50	0.55	41.48	0.21	17.03	0.93
Y61	22.23	0.68	29.61	0.68	33.22	0.23
Y62	*		31.37	0.69	43.80	**0.03**
Y63	31.28	0.45	39.93	0.16	29.47	0.49
Y64	36.28	0.17	24.72	0.85	30.77	0.38
Y65	31.32	0.35	33.87	0.43	32.59	0.30
Y66	29.03	0.41	31.44	0.69	20.44	0.88
Y67	37.86	0.13	31.64	0.68	23.35	0.67
Y68	*		47.45	**0.04**	29.27	0.45
Y69	35.10	0.24	29.25	0.74	27.18	0.51
Y70	32.14	0.36	53.10	**0.03**	23.91	0.69
Y71	29.49	0.39	18.80	0.17	27.82	0.53
Y72	28.21	0.51	28.66	0.59	27.37	0.60
Y73	27.73	0.43	35.84	0.25	29.07	0.57
Y74	27.83	0.53	50.69	0.05	22.99	0.69
Y75	35.83	0.18	*		38.62	0.13
Y76	29.63	0.38	35.68	0.30	31.02	0.47
Y77	13.29	0.99	28.49	0.77	37.15	0.21
Y78	29.05	0.36	28.70	0.73	32.08	0.23
Y79	22.62	0.48	35.33	0.50	38.85	0.13
Y80	36.86	0.10	24.52	0.82	53.36	**0.00**
Y81	*		27.44	0.65	*	

(Devam ediyor)

Tablo 14 (Devam)

ALES Sözel Alt Testlere Ait Madde-Model Uyum Değerleri

	Dönem 1		Dönem 2		Dönem 3	
Madde	S-χ2	p	S-χ2	p	S-χ2	p
Y82	*		37.01	0.42	22.33	0.77
Y83	28.01	0.41	36.05	0.17	39.28	0.08
Y84	21.20	0.85	38.51	0.23	32.74	0.29
Y85	29.58	0.49	31.23	0.70	27.22	0.45
Y86	22.05	0.82	25.89	0.81	47.22	**0.02**
Y87	*		30.11	0.70	35.86	0.21
Y88	42.48	0.05	58.26	**0.01**	36.71	0.15
Y89	32.72	0.34	38.72	0.23	19.64	0.77
Y90	23.91	0.81	37.45	0.27	38.43	0.06
Y91	24.30	0.76	41.02	0.13	29.37	0.34
Y92	*		46.01	0.08	45.36	**0.02**
Y93	31.39	0.26	37.82	0.22	26.52	0.44
Y94	26.76	0.59	33.52	0.49	37.17	**0.02**
Y95	36.91	0.12	37.93	0.30	52.89	**0.00**
Y96	26.09	0.51	40.47	0.24	32.28	0.12
Y97	114.97	**0.00**	43.00	0.11	27.22	0.51
Y98	55.08	**0.00**	78.06	**0.00**	38.66	0.09
Y99	100.29	**0.00**	62.15	**0.00**	30.40	0.45
Y100	24.01	0.58	55.50	**0.01**	28.15	0.40

Madde-model uyumu sağlamayan maddeler bold olarak gösterilmiştir.
* ile belirtilen maddeler DFA ve MTK'ya dayalı kestirimler sonucunda çıkartılan maddelerdir.

ALES'in sözel alt terslerinde yer alan 141 maddenin 3PL ile madde-model uyumunun incelenmesi amacıyla hesaplanan S-χ2 değerleri incelendiğinde 124 maddenin 3PL ile uyumlu olduğu ($p>0.05$), 17 maddenin ise madde-model uyumunu sağlamadığı ($p<0.05$) görülmüştür. ALES sözel alt testlerinde yer alan maddelerin %88'inin 3PL ile uyumlu olduğu sonucuna ulaşılmıştır.

Yapılan analizler sonucunda hem model-veri uyumu hem de madde-model uyumu incelendiğinde ALES alt testlerinin madde ve yetenek kestirimleri için 3PL modelinin uygun olduğu sonucuna ulaşılmıştır.

BOBUT Post-hoc Simülasyonları

ALES alt testleri için en uygun BOBUT koşullarının belirlenmesi amacıyla SimulCAT programı yardımıyla Post-hoc simülasyonları gerçekleştirilmiştir. SimulCAT, Chris Han tarafından 2010 yılında geliştirilmiş bir yazılımdır. Bu yazılım BOBUT

uygulamalarında Monte Carlo ve Post-hoc simülasyonları yapmak amacıyla tasarlanmıştır. BOBUT literatüründe üç temel araştırma deseni bulunmaktadır. Birincisi, Monte Carlo simülasyon çalışmasıdır; bu yöntem, belirli koşullar altında yanıt oluşturmak amacıyla birey ve madde parametrelerini simüle eder. İkincisi, post-hoc simülasyon çalışmasıdır ve bu yöntemde, BOBUT için gerçek bir madde havuzundan elde edilen madde parametreleri ve gerçek bireylerden kestirilen yetenek parametreleri kullanılır. Üçüncüsü ise, canlı BOBUT uygulamasıdır ve bu yöntem, gerçek adaylarla gerçek bir test ortamında uygulanır (Seo ve Choi, 2018). Post-hoc simülasyonları, bilgisayar ortamında belirli koşullar altında rastgele üretilen veriyi analiz eden Monte Carlo simülasyonlarının aksine önceden toplanmış gerçek veri üzerinde yapılan analizlerdir (Rubinstein ve Kroese, 2007). Gerçek veriye dayalı olarak yürütülen Post-hoc simülasyonları sonucunun ölçme kesinliği farklı istatistikler kullanılarak değerlendirilmiştir. BOBUT yaklaşımını ölçme kesinliği açısından karşılaştırmak için her bir koşula ait, kestirilen ve gerçek yetenek düzeyleri arasındaki uyum (fidelty), yanlılık (bias), ortalama standart hata (OSH) ve RMSE (root mean square error) incelenmiştir. Bu istatistikler ölçme kesinliğini değerlendirmek için yaygın kullanılmaktadır (Han, 2010; Seo ve Choi, 2018; Wang ve Vispoel 1998; Wang ve Zhang 2017). Uyum, kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki ilişkinin gücünü temsil eder ve bu ilişkinin yüksek olması beklenir. Yanlılık, sınava giren tüm bireyler için kestirilen yetenek düzeylerinin, gerçek yetenek düzeyleriyle arasındaki farkı temsil eder. RMSE ise, kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki farkların karelerinin ortalamasının karekökü̈dür (Tsaousis vd., 2021). Daha düşük bir yanlılık, OSH ve RMSE kestirilen yetenek düzeylerinin gerçek yetenek düzeylerine daha yakın olduğunu gösterir. Bu istatistiklerin hesaplaması şu şekildedir:

$$\text{Yanlılık} = \frac{\sum_{i=1}^{N}(\widehat{\theta}_i - \theta_i)}{N} \qquad \text{(Eşitlik 6)}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{\theta}_i - \theta_i)^2}{N}} \qquad \text{(Eşitlik 7)}$$

$$\text{OSH} = \frac{\sum_{i=1}^{N} SH(\theta_i)}{N} \qquad \text{(Eşitlik 8)}$$

Eşitlik 6, 7 ve 8'de verilen denklemlere göre "N" örneklem büyüklüğünü, θi i bireyinin gerçek yetenek düzeyini ve $\hat{\theta}_i$ ise i bireyinin kestirilen yetenek düzeyini tahmin eder. Bu istatistikler, SimulCAT post-hoc çıktıları kullanılarak Microsoft Excel programı yardımıyla hesaplanmıştır.

BOBUT yaklaşımının test güvenliği performansını değerlendirmek amacıyla, madde havuzunun kullanma oranları incelenmiştir. Bu analiz havuzda yer alan maddelerin ne kadar adil bir şekilde kullanıldığını ve hangi şartlarda havuzda yer alan maddelerin daha az kullanıldığını anlamaya yardımcı olur. Ayrıca, BOBUT yaklaşımının güvenilirliğini değerlendirmek için marjinal güvenilirlik katsayısı hesaplanmıştır. Marjinal güvenilirlik, testin güvenilirliğini ölçen bir istatistiktir ve standart hatanın bir fonksiyonu olarak hesaplanır. Marjinal güvenilirlik formülü şu şekildedir (Green vd., 1984):

$$MR = 1\text{-}OSH^2 \qquad \text{(Eşitlik 9)}$$

Post-hoc simülasyonları ile incelenen BOBUT koşulları aşağıda verilen Tablo 13'te yer almaktadır.

Tablo 15

BOBUT Post-hoc Koşulları

Koşul	Yetenek Kestirim Yöntemi	Madde Seçim Yöntemi	Test Sonlandırma Kuralı
Koşul 1	MO	MFB	SH<.30
Koşul 2	MO	KLB	SH<.30
Koşul 3	MO	AOB	SH<.30
Koşul 4	MO	VDB	SH<.30
Koşul 5	BSD	MFB	SH<.30
Koşul 6	BSD	KLB	SH<.30
Koşul 7	BSD	AOB	SH<.30
Koşul 8	BSD	VDB	SH<.30
Koşul 9	MO	MFB	SH<.40
Koşul 10	MO	KLB	SH<.40
Koşul 11	MO	AOB	SH<.40
Koşul 12	MO	VDB	SH<.40
Koşul 13	BSD	MFB	SH<.40
Koşul 14	BSD	KLB	SH<.40
Koşul 15	BSD	AOB	SH<.40
Koşul 16	BSD	VDB	SH<.40
Koşul 17	MO	MFB	SH<.50
Koşul 18	MO	KLB	SH<.50

(Devam ediyor)

Tablo 13 (Devam)

BOBUT Post-hoc Koşulları

Koşul	Yetenek Kestirim Yöntemi	Madde Seçim Yöntemi	Test Sonlandırma Kuralı
Koşul 19	MO	AOB	SH<.50
Koşul 20	MO	VDB	SH<.50
Koşul 21	BSD	MFB	SH<.50
Koşul 22	BSD	KLB	SH<.50
Koşul 23	BSD	AOB	SH<.50
Koşul 24	BSD	VDB	SH<.50
Koşul 25	MO	MFB	MS=15
Koşul 26	MO	KLB	MS=15
Koşul 27	MO	AOB	MS=15
Koşul 28	MO	VDB	MS=15
Koşul 29	BSD	MFB	MS=15
Koşul 30	BSD	KLB	MS=15
Koşul 31	BSD	AOB	MS=15
Koşul 32	BSD	VDB	MS=15
Koşul 33	MO	MFB	MS=20
Koşul 34	MO	KLB	MS=20
Koşul 35	MO	AOB	MS=20
Koşul 36	MO	VDB	MS=20
Koşul 37	BSD	MFB	MS=20
Koşul 38	BSD	KLB	MS=20
Koşul 39	BSD	AOB	MS=20
Koşul 40	BSD	VDB	MS=20

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi; SH: Standart Hata; MS: Madde Sayısı

Tablo 13'e göre ALES'in post hoc simülasyonu için her alt test için 40 farklı koşul incelenmiştir. Yetenek kestirim yöntemi olarak MO ve Bayesian kestirim yöntemi olan BSD yöntemleri kullanılmıştır. MO ve BSD yöntemleri BOBUT uygulamalarında sıklıkla kullanılmakta ve hangi yöntemin daha iyi sonuç verdiği literatürde tartışılan bir konudur. MO yönteminin Bayesian yetenek kestirim yöntemlerine göre yansız kestirim yaptığı ve büyük standart hata değeri ürettiği; Bayesian yöntemlerinin yanlı kestirim yaptığı ve Bayesian yöntemleri arasında ise BSD'nin göreceli olarak daha küçük standart hata ürettiği ve daha küçük yanlı kestirim yaptığı (Wang, 1997) bunun yanı sıra yapılan başka araştırmalarda MO yönteminin uyumsuz sonuçlar ürettiği BSD'nin ise üretmediği, BSD'nin MO'ya göre avantajlı olduğu (Chen ve Choi, 2009; Wang vd., 2012) sonucuna ulaşılmıştır. Bu karşıtlığın yanı sıra yapılan bazı araştırmalarda MO ve Bayes yetenek kestirim yöntemlerinin madde sayısı 15 ve üstü (Babcock ve Weiss, 2009), 20 ve üstü

(Thissen ve Mislevy, 2000) olduğunda benzer sonuçlar verdiği bulunmuştur. Bunun için bu araştırmada yetenek kestirim yöntemi olarak hem MO hem de BSD kullanılarak en uygun yöntem belirlenmiştir.

Bu araştırmada sabit ve değişen uzunluklu olmak üzere iki farklı test sonlandırma kuralı kullanılmıştır. Sabit uzunluk yönteminde testi alan her bireye aynı sayıda madde yönlendirilir. BOBUT uygulamaları sayesinde yetenek kestirimlerinde kullanılan madde sayısında %50 ile %80 arasında tasarruf sağlanmaktadır (Bulut ve Kan, 2012; Kalender, 2011; Kezer, 2013; Scullard, 2007; Smits vd, 2011). Bu oranlar dikkate alındığında ALES'in Türkçe ve Matematik alt testinin BOBUT uygulaması kapsamında, alt testte yer alan 50 maddeden %60 ve %70 oranında madde tasarrufu sağlayacak şekilde, sırasıyla 20 ve 15 maddeden oluşan iki farklı sabit test uzunluğu sonlandırma kuralının kullanılması planlanmıştır. Değişen uzunluklu test sonlandırma kuralı olan Standart hata (SH) yöntemi en yaygın kullanılan test sonlandırma kuralıdır (Babcock ve Weiss, 2012; Boyd vd., 2010). Bu yöntemde yetenek kestiriminin standart hatası önceden belirlenen sabit bir değere ulaştığında test sonlandırılmaktadır (Thompson ve Weiss, 2011). Alan yazın incelendiğinde bilişsel özellikleri ölçen BOBUT uygulamalarında (Bulut ve Kan, 2012; Kalender, 2011) .10 ve .50 arasındaki SH değerleri kullanıldığı görülmektedir. Bu nedenle bu araştırmada .30, .40 ve .50 SH sonlandırma yöntemleri kullanılmıştır. Bu değerler standart hata ve ölçme kesinliği arasındaki ilişki ($r = 1 - SH^2$) yardımıyla (Lau ve Wang, 1999) hesaplanan 0.91, 0.84 ve 0.75 marjinal güvenirliğe karşılık gelmektedir.

BOBUT post-hoc simülasyonlarında, testin sonlandırma kuralları büyük önem taşımaktadır. Özellikle çok kısa kalan BOBUT uygulamalarının, özellikle düşük yetenek düzeylerinde güvenilir kestirimler sağlayamadığı bilinmektedir (Babcock ve Weiss, 2012; van der Linden ve Pashley, 2010; Weiss, 1982). Ortalama test uzunluğu 10 maddenin altında olan uygulamalarda yanlılık ve RMSE gibi kestirim doğruluğu göstergelerinde anlamlı bozulmalar gözlemlenmiştir. Bununla birlikte, zaten uzun olan testlere (örneğin 50 madde) daha fazla madde eklenmesinin ise yetenek kestiriminin doğruluğunda anlamlı bir artış sağlamadığı belirtilmiştir (Babcock ve Weiss, 2012). Bu sebeple, 1-0 (dikotom) olarak puanlanan maddelerle çalışan BOBUT uygulamalarında, kestirim doğruluğunu artırmak amacıyla en az 10–15 madde sunulması önerilmektedir (Babcock ve Weiss, 2012; Bock ve Mislevy, 1982). Ayrıca SH değerinin BOBUT uygulaması sonlandırmada tek başına kullanılması durumunda, özellikle test çok kısa olduğunda ya da testin uzunluğu değişken olduğunda, ölçme hassasiyeti üzerinde yeterli düzeyde kontrol sağlanamayabileceği belirlenmiştir (Lee ve Han, 2024). Bu gerekçelere

dayanarak, ALES alt testlerinin her birinde 50 madde bulunması ve bu sınavın yüksek riskli bir seçme sınavı olması dikkate alınarak, post-hoc BOBUT simülasyonlarında SH sonlandırma koşulları için minimum madde sayısı 10, maksimum madde sayısı ise 50 (alt test uzunluğu) olarak belirlenmiştir. ALES post-hoc simülasyonlarında minimum 10 madde sınırlandırılması yapılmadığında bazı bireylerde yalnızca 1 maddeyle yetenek kestirimi yapılabildiği gözlemlenmiş, ancak bu tür durumlar, kamuoyunda sınavın güvenilirliğine yönelik olumsuz algıların oluşmasına neden olabileceği için minimum 10 madde kullanılmıştır.

Bu araştırmada madde seçim yöntemi olarak MFB, KLB, AOB ve VDB yöntemleri kullanılmıştır. Madde seçim yöntemlerinde, geçici yetenek tahmininde maksimum bilgi sağlayan maddeyi seçmeye yönelik olan madde bilgisine dayalı yaklaşımlar bulunmaktadır (Choi ve Swartz, 2009; van der Linden, 1998; van der Linden ve Pasley, 2000). Bu araştırmada incelemek için seçilen MFB, KLB ve AOB yöntemleri madde bilgisine dayalı yöntemlerdir. MFB yöntemi en çok kullanılan madde seçim yöntemi (Lord, 1977) olmasına rağmen bireyin yetenek düzeyi hakkında en yüksek bilgiyi verecek maddelerin sıklıkla kullanması ve bireyin yetenek düzeyine uygun ama daha az ayırıcı olan bazı maddeleri nadiren kullanması veya hiç kullanmaması gibi bazı dezavantajlarından (Chang, 2014) dolayı bir diğer madde seçme yöntemi olan KLB yöntemi de araştırılmıştır. KLB yöntemi MFB yönteminin bir alternatifi olarak Chang ve Ying (1996) tarafından geliştirilmiştir. MFB yerel bilgi olarak adlandırılan küçük bir θ bölgesinin etrafındaki bilgiyi temsil ederken, KLB ise global bilgi olarak adlandırılan ve bölgenin dışındaki bilgiyi temsil eder, bu da θ seviyelerinin daha geniş bir ranjda farklılaşma gücünü gösterir. AOB yöntemi Veerkamp ve Berger (1997) tarafından MFB'ye alternatif olarak geliştirilmiştir. Bu yöntem, θ düzeyindeki belirsizlikleri dikkate alarak daha genel ve kararlı madde seçimleri yapmayı hedefler. Madde seçiminde belirsizliği daha iyi yönetmek için, sadece tek bir yetenek tahmini noktasındaki bilgiye bakılmaz. Bunun yerine, kişinin yetenek seviyesinin farklı olabileceği tüm olası değerler göz önünde bulundurulur (Han, 2012). Bu araştırmada ayrıca VDB yöntemi kullanılmıştır. Bu yöntem ise BOBUT uygulamalarının erken aşamalarında daha düşük ayırt ediciliğe sahip maddeleri kullanılırken, bireyin yetenek düzeyi netleştikçe yüksek ayırt edicilik değerine sahip maddeleri tercih eder. Bu durum ise θ noktasında en yüksek bilgiyi veren maddeyi kullanmaya odaklı olmayıp θ aralığı boyunca değerlendirilir (Cheng vd., 2008; Han, 2012). Sonlandırma kuralı, yetenek kestirim ve madde seçim yönteminin manipüle edildiği çalışmada; ilk maddenin seçimi: θ yetenek düzeyi (-0.5 ile

0.5 arasında), madde kullanım sıklığı kontrolü Randomesque, kapsam dengeleme ise Kısıtlanmış BOBUT (constrained CAT) olacak şekilde sabit tutulmuştur. İçerik dengelemesi için sayısal ve sözel alt testlerde yer alan maddelerin içerik dağılımları uzman görüşleri yardımıyla incelenmiş olup Tablo 14'te verilmiştir.

Tablo 16

Sayısal ve Sözel Alt Testin İçerik Dağılımı (3 Form İçin)

Alt Test / İçerik	**Toplam madde sayısı**	**Yüzde (%)**
SAYISAL		
Sayılarla İşlem	45	30.0
Sayısal mantık ve problem çözme	84	56.0
Geometri	21	14.0
Sayısal Alt Test Toplam	**150**	**100.0**
SÖZEL		
Metin yapısı ve anlam bütünlüğü	51	36.0
Paragraf	66	47.0
Sözel mantık	24	17.0
Sözel Alt Test Toplam	**141**	**100.0**

ÖSYM'ye göre (2018a) ALES'in sayısal testi, sayısal ve mantıksal akıl yürütme(muhakeme); sözel testi ise sözel akıl yürütme becerilerini ölçmeye yönelik çoktan seçmeli maddelerden oluşmaktadır. Dolayısıyla her alt testteki tüm maddeler aynı temel beceriyi değerlendirmektedir. Ancak maddeler içerik bağlamı bakımından çeşitlilik gösterdiğinden, bu çalışmada içerik dengelemesi uygulanarak adaylar için daha adil ve eşdeğer test koşulların sağlanması hedeflenmiştir. Böylelikle, sınavın gerçek içerik dağılımı (alanlar ve madde oranları) korunmakta; simülasyon gerçek sınav koşullarına daha yakın hale gelmekte ve dolayısıyla daha güvenilir sonuçlar elde edilmektedir. Alt testlerin içerik incelenmesi Türkçe, Matematik ve ölçme değerlendirme uzmanları tarafından değerlendirilmiştir. Matematik alt testinde yer alan rasyonel sayı, denklem çözme, üslü sayılar gibi maddeler "sayılarla işlem" alt içeriği altında, problem çözme ve mantıksal çıkarımların yer aldığı maddeler "sayısal mantık ve problem çözme" alt içeriği altında, testin sonunda yer alan geometri maddeleri ise "geometri" alt içeriği altında toplanmıştır. Buna paralel olarak, sözel alt testte yer alan cümle/paragraf tamamlama, cümle sıralama ve anlatım bozukluğunu tespit etme gibi maddeler "Metin Yapısı ve Anlam Bütünlüğü" alt başlığı altında; paragraf okuma ve paragraftan anlam çıkarma ile ilgili maddeler "Paragraf" alt başlığı altında; testin sonunda yer alan sözel mantık soruları

ise “Sözel Mantık” alt başlığı altında toplanmıştır. İçerikte yer alan madde oranları ise, üç formdaki toplam 150 sayısal ve 141 sözel maddeden türetilmiş, SimulCAT’te *Content Balancing› By Weight* seçeneğine aynı yüzdeler girilerek uygulanmıştır. Post hoc simülasyonları, sayısal alt testte 2.127 birey ve 150 maddeden oluşan bir madde havuzu ile; sözel alt testte ise 4.323 birey ve 141 maddeden oluşan bir madde havuzu kullanılarak gerçekleştirilmiştir.

ALES Canlı BOBUT Uygulamasının Geliştirilmesi

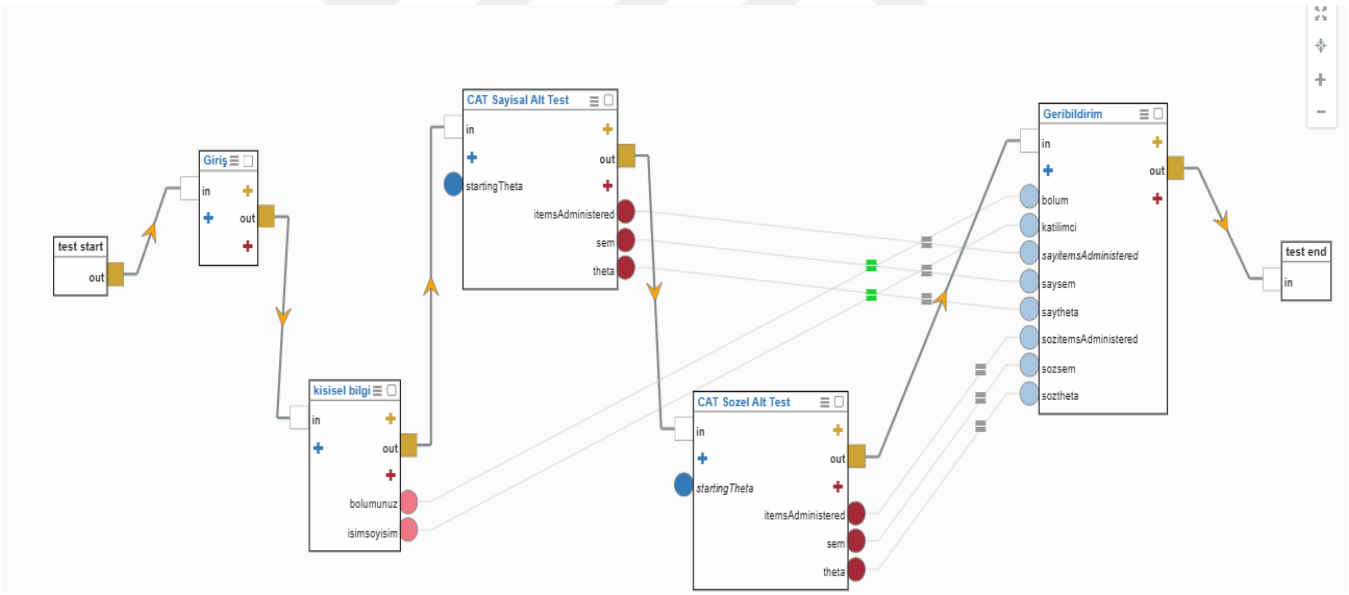
Araştırmanın ikinci aşaması olan ALES canlı BOBUT uygulaması Cambridge Üniversitesi Psikometri Merkezi tarafından geliştirilen, açık kaynaklı ücretsiz bir yazılım olan Concerto platformu yardımıyla yapılmıştır. Concerto, özellikle bireyselleştirilmiş/uyarlanabilir (adaptive) testler geliştirmek ve dağıtmak için kullanılan arka planda R programlama dili kullanan açık kaynaklı bir bilgisayar tabanlı test platformudur. Concerto Platformu’nu kullanabilmek için, https://concertoplatform.com/about adresindeki GitHub sayfası üzerinden platform dosyalarına erişmek ve ardından Amazon Web Services (AWS) gibi bir bulut hizmeti kullanarak sunucu kurulumunu gerçekleştirmek gerekmektedir. Kurulum sırasında admin kullanıcı adı ve şifresi belirlenir. Bu kullanıcı adı ve şifresini kullanarak bu platform aracılığıyla BOBUT uygulaması gerçekleştirilebilmektedir. Concerto’da her BOBUT, bir akış (flow) yapısı üzerinden ilerler. Her adım, bir "node" olarak tanımlanır. Concerto Platformu, farklı kullanım amaçlarına hizmet eden çok sayıda node yapısı içermektedir. Bu çalışmanın amacı doğrultusunda, alt testlerin ayrı ayrı uygulanabilmesi için “CAT sayısal alt test” ve “CAT sözel alt test” olmak üzere iki ayrı "Assessment" node’u kullanılmış ve bu node’lar test akışı içerisinde birbirine bağlanmıştır. Test node akışı şekil 7’de gösterilmiştir. Sayısal alt testte yer alan 150 madde ile sözel alt testte yer alan 141 madde, ekran görüntüsü alınarak Concerto platformuna ayrı dosyalar hâlinde yüklenmiştir. Alt testlerdeki maddelerin şekilsel, formülsel ve yazınsal açıdan yüksek düzeyde yoğunluk içermesi, bu maddelerin tek tek metin formatında yazılmasını güçleştirmektedir. Bu nedenle, kullanım kolaylığı sağlamak ve süreci daha verimli hâle getirmek amacıyla maddeler ekran görüntüsü formatında sisteme aktarılmıştır. Alt testlere ilişkin test akışı, post hoc simülasyonlar sonucunda elde edilen en uygun BOBUT koşulları temel alınarak yapılandırılmıştır. ALES canlı BOBUT uygulamasında, madde seçim yöntemi olarak MFB (MFI); θ kestirimi için BSD (EAP) ve sonlandırma kuralı

olarak MS=20 sabit madde sayılı durdurma kuralı kullanılmıştır. Test başlatma kuralı olarak, θ = -0.5 – 0.5 aralığı kullanılmıştır. Her madde için 3 dakika yanıtlama süresinin uygun olacağı düşünülmüş ve katılımcının her bir alt testte çözeceği maksimum 20 madde için 60 dk süre verilmiştir. Canlı BOBUT uygulamasına madde kullanım sıklığı ve kapsam dengeleme koşulları da eklenmiştir. Madde kullanım sıklığını dengelemek amacıyla, her uygulama adımında bireyin kestirilen geçici yetenek düzeyine en uygun beş madde arasından rastgele seçim yapılması yoluna gidilmiştir. Ayrıca, her bir alt testte kapsam dengelemesini sağlamak amacıyla, her bir konu alanının (sayısal alt test için; sayılarla işlem (%30), sayısal mantık ve problem çözme (%56), geometri (%14); sözel alt test için; dil bilgisi ve anlam bütünlüğü (%36), paragraf (%47), sözel mantık (%17) testte temsil edilme oranları yüzdelik değerler olarak sisteme tanımlanmıştır.

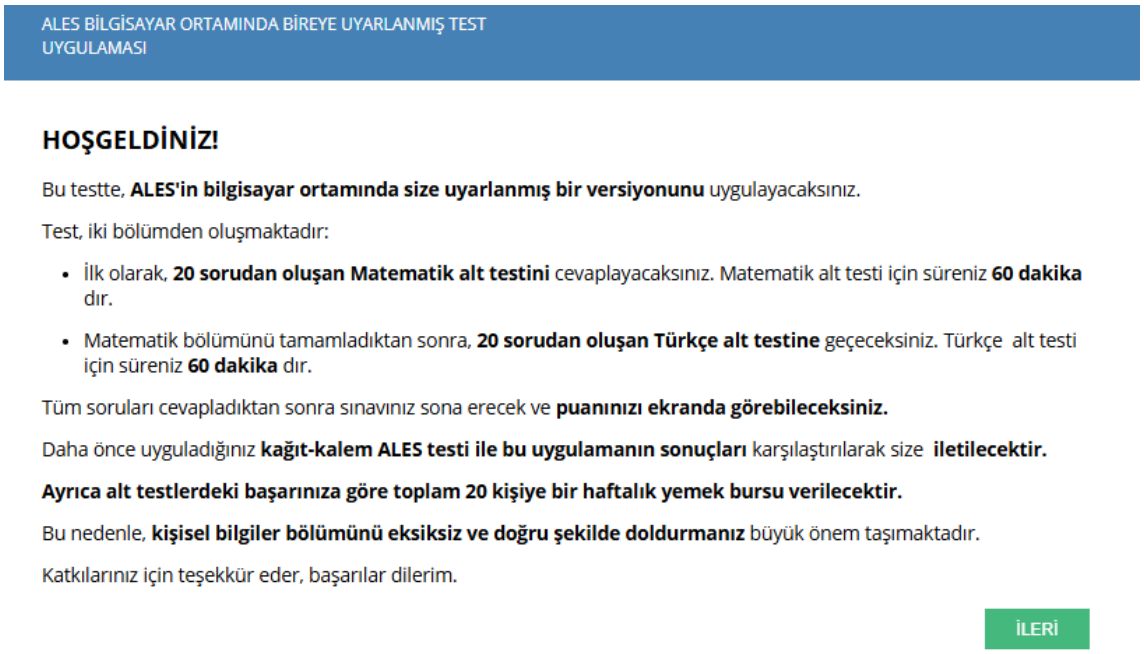
ALES BOBUT uygulaması Adıyaman Üniversitesi fakültelerinde yer alan bilgisayar laboratuvarlarında; ALES kağıt-kalem testleri ise ders saatleri içerisinde fakültedeki dersliklerde gerçekleştirilmiştir. Bu süreçte üniversitenin mevcut bilgisayar altyapısı ve internet ağı kullanılmıştır. Tüm uygulamalar araştırmacının gözetiminde ve doğrudan katılımıyla yürütülmüş; uygulama sırasında herhangi bir teknik ya da yönetsel sorun yaşanmaması için araştırmacı aktif olarak görev almıştır. Uygulamalar, aralarında ortalama bir haftalık bir süre olacak şekilde; bazı gruplara önce ALES kağıt-kalem testi, bazı gruplara ise önce ALES canlı BOBUT uygulaması yapılacak biçimde planlanmış ve uygulanmıştır.

Katılımcılar ALES canlı BOBUT uygulamasına https://eptlab.com/ales adresinden erişim sağlamıştır. Şekil 8'de verilen ilk karşılama ekranındaki yönergeyi okuduktan sonra ad-soyadı, bölüm gibi kişisel bilgilerini girerek uygulamaya başlamışlardır. Katılımcılar, testin ilk aşamasında ALES'in kâğıt-kalem uygulamasına benzer biçimde yapılandırılmış, 20 maddeden oluşan bir matematik alt testini tamamlamışlardır. Matematik testi tamamlandıktan sonra, ikinci aşamada yine 20 maddeden oluşan bir Türkçe alt testi uygulanmıştır. Testin sonunda ise, bireylere her iki alt test için ayrı ayrı yetenek düzeyi (theta), standart hata (SH) ve uygulanan madde numaraları gibi bilgileri içeren bir geribildirim ekranı sunulmuştur. Bu geribildirim ekranı Şekil 9'da gösterilmiştir.

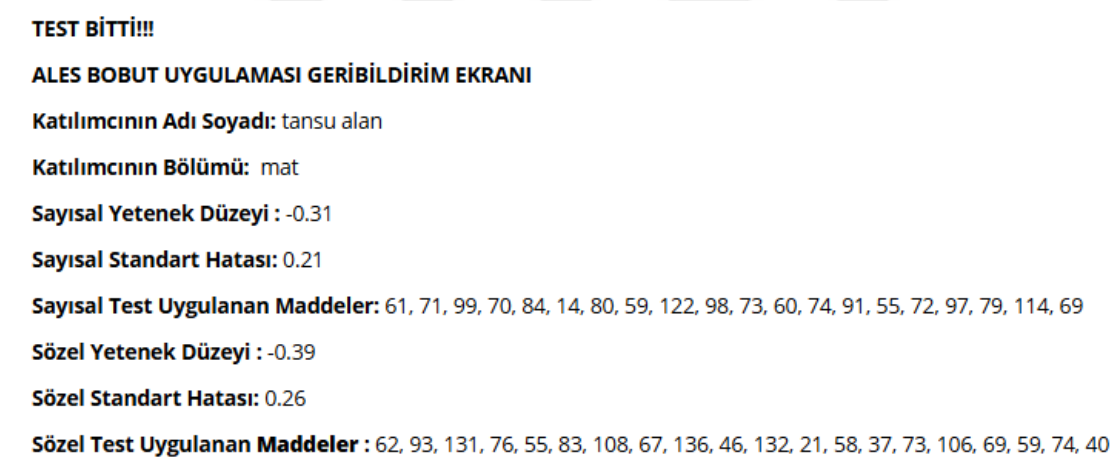
Ayrıca çalışmada, katılımcıları ALES çözmeye teşvik etmek, motive etmek ve sınav sürecine aktif katılımlarını artırmak amacıyla bir ödül sistemi uygulanmıştır. Ödül sistemi, ALES kâğıt-kalem uygulaması ile ALES canlı BOBUT uygulamasından ayrı ayrı elde edilen yetenek kestirimlerine (θ) dayanmaktadır. Kâğıt-kalem ve canlı BOBUT uygulamalarında, her biri sayısal ve sözel alt testlerden oluşan θ değerleri ayrı ayrı hesaplanmıştır. Bu kapsamda, iki uygulamada ayrı ayrı her alt test için θ değeri en yüksek 5 katılımcı belirlenmiş ve toplamda 20 bursiyer olacak şekilde bir haftalık yemek bursu verilmesi planlanmıştır. Katılımcılara uygulamalar öncesinde ödül sistemi hakkında bilgi verilmiştir. Aynı katılımcının birden fazla listede yer alması durumunda, ilgili katılımcıya burs yalnızca bir kez tahsis edilir; toplam bursiyer sayısını 20'de tutmak için, yer aldığı her listede sıradaki öğrenci (6., gerekirse 7. vb.) listeye dâhil edilir. Yetenek düzeylerinin eşitliği hâlinde öncelik sırası: (i) θ kestirimlerinin standart hatası (SH) daha düşük olan, (ii) SH'de eşitse alt testteki toplam doğru sayısı daha yüksek olan şeklinde belirlenmiştir. Böylece her iki uygulama ve alt testlerden elde edilen performanslar dengeli biçimde ödüllendirilirken, toplam kontenjan sabit tutulmuştur.



Şekil 7. ALES Canlı BOBUT Akışı



Şekil 8. ALES Canlı BOBUT Karşılama Ekranı



Şekil 9. ALES Canlı BOBUT Geribildirim Ekranı

BÖLÜM 4

BULGULAR VE YORUMLAR

Bu bölümde, araştırmanın alt amaçlarına yönelik elde edilen bulgular ilgili alt başlıklar altında sistematik bir şekilde sunulmuştur.

ALES Madde Havuzunda Yer Alan Maddelerin MTK'ya Göre Parametreleri

Araştırmanın birinci alt amacına doğrultusunda 2021 üç dönem ALES sayısal ve sözel alt testlerine ait madde parametreleri model veri uyumu doğrultusunda 3PL modele göre kestirilmiştir. ALES sayısal alt testlerinde bulunan maddelere ilişkin hesaplanan güçlük, ayırt edicilik ve şans parametreleri Tablo 15'te sunulmuştur.

Tablo 17

ALES Dönem 1 Sayısal Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y1	1.190	-2.549	0.001	Y26	0.831	-3.004	0.003
Y2	1.693	-2.312	0.002	Y27	1.051	-1.395	0.002
Y3	1.383	-1.881	0.006	Y28	1.603	-0.505	0.150
Y4	2.471	-2.123	0.000	Y29	1.987	-1.209	0.244
Y5	0.918	-3.272	0.009	Y30	1.456	-1.675	0.002
Y6	1.345	-0.857	0.162	Y31	2.545	-1.571	0.137
Y7	1.795	-2.214	0.003	Y32	2.185	-1.405	0.106
Y8	0.746	-1.913	0.002	Y33	1.737	-0.371	0.164
Y9	1.355	-1.060	0.182	Y34	1.836	-1.258	0.000
Y10	2.128	-1.596	0.286	Y35	1.651	-1.334	0.100
Y11	2.201	-1.920	0.005	Y36	1.635	-0.536	0.167
Y12	1.367	-1.940	0.001	Y37	1.558	-1.354	0.001
Y13	1.967	-1.562	0.339	Y38	1.689	-1.824	0.001
Y14	0.747	-1.026	0.002	Y39	1.326	-1.562	0.074
Y15	1.997	-1.781	0.116	Y40	0.937	-0.388	0.167
Y16	1.589	-1.939	0.001	Y41	1.978	-2.020	0.000
Y17	1.524	-2.372	0.003	Y42	1.612	-1.578	0.003
Y18	2.632	-1.252	0.221	Y43	1.396	-1.776	0.002
Y19	1.385	-0.933	0.049	Y44	1.712	-1.701	0.058
Y20	2.712	-1.241	0.208	Y45	2.197	-1.261	0.171
Y21	1.245	-1.612	0.194	Y46	1.152	-1.656	0.012
Y22	1.821	-1.232	0.109	Y47	1.596	-1.015	0.070
Y23	1.193	-0.672	0.091	Y48	1.672	-1.301	0.103
Y24	0.985	-0.604	0.007	Y49	0.843	-1.893	0.002
Y25	0.481	-0.542	0.015	Y50	1.043	-1.747	0.001

Tablo 15 incelendiğinde dönem 1 sayısal alt testinde yer alan maddeler incelendiğinde, b parametrelerinin-3.272 ile -0.371 arasında yer aldığı ve testin kolay maddelerden oluştuğu söylenilebilir. Y25 maddesinin a parametresi 0.64'ün altında olmasından dolayı düşük ayırt ediciliğe sahip olduğu sonucuna ulaşılmıştır. Alt testteki diğer 49 maddenin ayırt edicilik parametrelerinin 0.65'in üzerinde olması, maddelerin ayırt edicilik bakımından orta ve yüksek düzeyde olduğunu göstermektedir. Şans parametreleri ise tüm maddelerde kabul edilebilir sınır içinde yer almasından dolayı dönem 1 sayısal alt testte yer alan tüm maddelerin post-hoc simülasyonlarına alınmasına karar verilmiştir.

Tablo 18

ALES Dönem 2 Sayısal Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y1	1.992	-1.632	0.001	Y26	2.366	0.337	0.183
Y2	2.276	-1.166	0.010	Y27	1.851	-0.983	0.005
Y3	2.694	-0.605	0.071	Y28	1.578	-0.824	0.000
Y4	2.448	-1.162	0.035	Y29	2.501	-0.555	0.083
Y5	2.773	-0.645	0.079	Y30	2.625	0.579	0.181
Y6	1.788	-1.328	0.001	Y31	3.531	0.010	0.061
Y7	2.430	-1.216	0.001	Y32	1.690	-0.555	0.109
Y8	1.889	-0.637	0.001	Y33	1.638	-0.284	0.066
Y9	1.173	-0.525	0.000	Y34	1.529	1.169	0.181
Y10	2.325	-0.377	0.107	Y35	2.061	-1.381	0.002
Y11	3.803	0.379	0.121	Y36	1.768	-0.388	0.003
Y12	2.742	0.105	0.163	Y37	1.496	-1.158	0.001
Y13	3.798	-0.808	0.152	Y38	2.361	-1.055	0.064
Y14	1.700	-1.090	0.072	Y39	2.706	-1.000	0.039
Y15	2.104	-1.185	0.000	Y40	1.601	1.251	0.128
Y16	2.489	-0.393	0.129	Y41	2.122	-0.407	0.101
Y17	2.623	-1.002	0.371	Y42	1.278	-0.810	0.001
Y18	2.132	-0.418	0.024	Y43	2.896	-0.666	0.000
Y19	1.749	-0.547	0.000	Y44	2.668	-0.928	0.048
Y20	4.835	0.491	0.252	Y45	2.535	-0.896	0.070
Y21	1.712	0.121	0.104	Y46	3.679	0.078	0.145
Y22	2.262	-0.289	0.165	Y47	2.607	-0.696	0.071
Y23	2.612	-0.083	0.158	Y48	1.935	-0.070	0.137
Y24	3.179	-0.188	0.132	Y49	1.187	-0.468	0.085
Y25	2.946	0.239	0.253	Y50	1.000	-0.377	0.001

Tablo 16 incelendiğinde, dönem 2 sayısal alt testinde yer alan maddeler incelendiğinde, b parametrelerinin -1.632 ile 1.251 arasında yer aldığı ve testin kolay,

orta ve zor maddelerden oluştuğu söylenilebilir. Testte yer alan maddelerin a parametresi 0.65'in üzerinde olması, maddelerin ayırt edicilik bakımından orta ve yüksek düzeyde olduğunu göstermektedir. Şans parametreleri incelendiğinde ise Y17 nolu maddenin 0.371 ile kabul sınırı civarında; kalan 49 maddenin ise kabul edilebilir sınır içinde yer almasından dolayı dönem 2 sayısal alt testte yer alan tüm maddelerin post-hoc simülasyonlarına alınmasına karar verilmiştir.

Tablo 19

ALES Dönem 3 Sayısal Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y1	1.421	-2.307	0.356	Y26	1.312	-2.286	0.001
Y2	1.995	-1.881	0.000	Y27	1.789	-2.202	0.001
Y3	3.546	-1.747	0.341	Y28	1.420	-1.854	0.112
Y4	1.437	-2.202	0.001	Y29	2.790	-1.514	0.359
Y5	1.171	-2.267	0.001	Y30	2.411	-1.175	0.189
Y6	1.519	0.056	0.348	Y31	1.198	-2.218	0.001
Y7	2.455	-2.169	0.000	Y32	1.857	-1.987	0.053
Y8	1.054	-0.946	0.193	Y33	3.414	-1.291	0.229
Y9	1.591	-1.976	0.001	Y34	1.559	-1.995	0.003
Y10	2.191	-1.627	0.254	Y35	1.227	-1.714	0.001
Y11	2.944	-1.715	0.186	Y36	2.570	-1.477	0.164
Y12	2.099	-2.008	0.009	Y37	3.155	-0.880	0.219
Y13	1.652	-1.300	0.073	Y38	0.817	-0.401	0.088
Y14	1.370	-0.788	0.190	Y39	0.921	-1.983	0.055
Y15	1.550	-1.929	0.001	Y40	1.426	-1.231	0.157
Y16	1.806	-1.416	0.090	Y41	1.226	-1.832	0.000
Y17	2.177	-0.811	0.254	Y42	1.365	-1.410	0.068
Y18	1.772	-1.857	0.000	Y43	1.738	-0.694	0.157
Y19	1.598	-1.228	0.121	Y44	1.722	-1.745	0.000
Y20	3.408	-2.132	0.004	Y45	1.934	-1.299	0.074
Y21	1.091	-0.428	0.009	Y46	1.898	-1.025	0.071
Y22	1.255	-0.268	0.088	Y47	1.902	-0.865	0.069
Y23	1.702	-2.288	0.068	Y48	2.143	-1.091	0.163
Y24	2.220	-1.439	0.162	Y49	1.237	-1.425	0.090
Y25	0.678	-0.697	0.001	Y50	1.557	-2.044	0.001

Tablo 17 incelendiğinde, dönem 3 sayısal alt testinde yer alan maddeler incelendiğinde, b parametrelerinin -2.307 ile 0.056 arasında yer aldığı ve testin kolay ve orta düzey zorlukta maddelerden oluştuğu söylenilebilir. Testte yer alan maddelerin a parametresi 0.65'in üzerinde olması, maddelerin ayırt edicilik bakımından yeterli düzeyde olduğunu göstermektedir. Şans parametreleri incelendiğinde ise tüm maddelerin

kabul edilebilir sınır içinde yer almasından dolayı dönem 3 sayısal alt testte yer alan tüm maddelerin post-hoc simülasyonlarına alınmasına karar verilmiştir.

Tablo 20

ALES Dönem 1 Sözel Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y51	0.531	2.209	0.373	Y76	0.543	0.629	0.001
Y52	0.548	3.114	0.176	Y77	1.016	-2.569	0.006
Y53	2.052	2.335	0.212	Y78	1.190	-2.396	0.002
Y54	0.998	1.529	0.523	Y79	1.469	-2.729	0.004
Y55	0.656	-2.383	0.004	Y80	1.077	-1.489	0.003
Y56	0.731	0.245	0.219	*Y81	0.340	1.093	0.005
Y57	0.729	-2.304	0.011	*Y82	0.077	15.891	0.038
Y58	1.104	-2.196	0.004	Y83	1.292	0.485	0.071
Y59	0.948	-1.744	0.003	Y84	1.215	-1.447	0.196
Y60	0.697	-0.986	0.032	Y85	1.079	-0.337	0.265
Y61	1.400	-1.888	0.030	Y86	0.835	-2.258	0.003
*Y62	-0.284	-4.848	0.002	*Y87	0.148	-2.951	0.021
Y63	0.388	-0.021	0.068	Y88	0.963	-1.728	0.001
Y64	0.672	0.661	0.099	Y89	0.813	-2.025	0.002
Y65	0.368	2.709	0.003	Y90	0.384	-1.100	0.004
Y66	0.866	-1.278	0.029	Y91	0.668	-1.484	0.007
Y67	0.918	0.294	0.159	*Y92	0.262	-2.443	0.015
*Y68	0.350	-0.865	0.010	Y93	1.512	-0.580	0.177
Y69	0.404	-3.711	0.008	Y94	1.404	-0.712	0.270
Y70	0.548	-2.079	0.008	Y95	1.339	-0.168	0.178
Y71	0.938	-0.660	0.167	Y96	1.849	-0.302	0.212
Y72	0.518	-0.710	0.003	Y97	8.105	0.610	0.172
Y73	1.354	-1.303	0.001	Y98	1.564	-0.425	0.150
Y74	1.001	-2.372	0.002	Y99	8.309	0.584	0.21
Y75	0.590	0.145	0.046	Y100	2.562	-0.352	0.295

*Testten çıkarılan maddeleri temsil etmektedir.

Tablo 18 incelendiğinde, dönem 1 sözel alt testinde yer alan maddeler incelendiğinde, b parametrelerinin-3.711 ile 3.114 arasında yer aldığı ve testin kolay, orta ve zor maddelerden oluştuğu söylenilebilir. Testte yer alan Y62, Y68, Y81, Y82, Y87 ve Y92 maddeleri (MTK varsayımlarının incelenmesi başlığı altında yürütülen DFA sonuçları ile birlikte incelenerek karar verilmiştir) ayırt edicilik bakımından çok düşük olduğundan testten çıkarılmıştır. Şans parametreleri incelendiğinde, Y54 maddesinin kabul edilebilir sınırların dışında olduğu görülmüştür. Ancak ayırt edicilik düzeyinin yüksek olması nedeniyle testte yer almasına karar verilmiştir. Diğer tüm maddeler ise kabul edilebilir sınırlar içerisinde yer almaktadır. Kalan 44 maddenin ise orta ve yüksek

düzeyde ayırt ediciliğe sahip olmasından dolayı post-hoc simülasyonlarına alınmasına karar verilmiştir.

Tablo 21

ALES Dönem 2 Sözel Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y51	0.585	-1.731	0.006	Y76	0.856	-0.504	0.002
Y52	1.113	-0.699	0.002	Y77	0.909	-1.014	0.021
Y53	0.745	0.581	0.003	Y78	0.774	-0.777	0.002
Y54	0.760	1.243	0.148	Y79	0.727	-1.071	0.002
Y55	0.638	-2.042	0.004	Y80	0.937	-0.969	0.002
Y56	0.937	-1.208	0.096	Y81	1.210	-1.085	0.009
Y57	0.640	-0.067	0.003	Y82	0.613	-1.272	0.002
Y58	0.988	-1.652	0.005	Y83	1.772	-0.892	0.105
Y59	0.850	-1.649	0.002	Y84	0.583	1.117	0.003
Y60	0.764	-1.511	0.011	Y85	0.645	-0.894	0.001
Y61	1.189	-1.534	0.045	Y86	1.221	-1.732	0.005
Y62	0.820	-1.438	0.154	Y87	0.869	-1.294	0.001
Y63	1.008	-0.258	0.066	Y88	0.795	-2.118	0.003
Y64	0.842	-0.825	0.002	Y89	1.152	-1.864	0.007
Y65	1.277	-1.627	0.031	Y90	1.268	-1.236	0.034
Y66	0.691	-2.249	0.017	Y91	1.223	-2.055	0.005
Y67	0.546	0.196	0.070	Y92	1.169	-1.456	0.001
Y68	0.955	-0.040	0.021	Y93	2.753	0.512	0.163
Y69	0.859	-1.168	0.003	Y94	2.484	0.418	0.280
Y70	0.721	-2.309	0.008	Y95	2.053	0.191	0.280
Y71	1.794	-3.003	0.024	Y96	2.145	-0.035	0.414
Y72	1.179	-0.881	0.005	Y97	4.557	0.389	0.298
Y73	1.366	-0.722	0.000	Y98	4.229	0.453	0.356
Y74	0.553	-1.018	0.004	Y99	3.306	0.654	0.163
*Y75	-0.190	-4.060	0.009	Y100	6.080	0.459	0.279

*Testten çıkarılan maddeleri temsil etmektedir.

Tablo 19 incelendiğinde, dönem 2 sözel alt testinde yer alan maddelerin b parametrelerinin -3.000 ile 1.243 arasında yer aldığı ve testin kolay, orta ve zor maddelerden oluştuğu söylenilebilir. Testte yer alan Y96 maddesinin c parametresinin kabul sınırları dışında olduğu ama a parametresinin yüksek olmasından dolayı testte yer alması kararlaştırılmıştır. MTK'ya dayalı kestirilen madde ayırt edicilik katsayısının pozitif olması beklenir (Baker, 2001). Bu sebeple negatif ayırt ediciliğe sahip Y75 maddesi ayırt edicilik bakımından yeterli düzeyde olmadığından testten çıkarılmıştır. Kalan 49 maddenin ise orta ve yüksek düzeyde ayırt ediciliğe sahip olduğu ve c

parametrelerinin kabul sınırları içinde olmasından dolayı post-hoc simülasyonlarına alınmasına karar verilmiştir.

Tablo 22

ALES Dönem 3 Sözel Alt Testi Madde Parametreleri

Madde No	a	b	c	Madde No	a	b	c
Y51	1.101	0.143	0.665	Y76	1.127	0.805	0.406
Y52	0.398	-2.288	0.102	Y77	0.549	-1.383	0.011
Y53	1.008	-0.664	0.655	Y78	1.061	-3.019	0.006
Y54	1.080	-2.130	0.538	Y79	0.738	-2.183	0.009
*Y55	0.316	-7.307	0.100	Y80	0.808	-2.965	0.007
Y56	1.060	-0.388	0.393	*Y81	1.460	3.130	0.048
Y57	1.451	-1.536	0.578	Y82	1.349	-1.924	0.375
Y58	0.903	-0.609	0.300	Y83	1.114	-2.648	0.009
Y59	0.445	-2.012	0.346	Y84	1.037	-2.039	0.090
Y60	1.546	1.025	0.137	Y85	1.278	-2.540	0.004
Y61	0.609	-0.465	0.025	Y86	0.990	-0.725	0.511
Y62	1.461	-1.347	0.398	Y87	0.536	-2.564	0.015
Y63	0.675	-2.412	0.011	Y88	1.145	-2.305	0.003
Y64	1.123	-0.032	0.272	Y89	1.432	-1.660	0.015
Y65	0.647	-1.815	0.499	Y90	2.091	-0.132	0.212
Y66	1.420	-1.004	0.343	Y91	1.510	-0.817	0.236
Y67	1.375	1.272	0.138	Y92	0.998	-1.256	0.002
Y68	1.291	0.445	0.319	Y93	1.405	-1.280	0.039
Y69	1.094	0.870	0.205	Y94	2.510	-1.298	0.055
Y70	0.892	0.664	0.161	Y95	1.903	-0.863	0.001
Y71	1.278	-1.361	0.332	Y96	1.557	-0.585	0.015
Y72	1.250	-0.670	0.339	Y97	1.331	-2.073	0.009
Y73	0.421	1.069	0.393	Y98	1.351	-1.924	0.005
Y74	1.151	-1.828	0.535	Y99	0.969	-1.687	0.001
Y75	0.780	-2.419	0.035	Y100	1.560	-0.182	0.159

*Testten çıkarılan maddeleri temsil etmektedir.

Tablo 20 incelendiğinde, dönem 3 sözel alt testinde yer alan maddelerin b parametrelerinin -3.019 ile 3.130 arasında yer aldığı ve testin kolay, orta ve zor maddelerden oluştuğu söylenilebilir. Y55 maddesinin ayırt ediciliğinin çok düşük, Y81 maddesinin ise anlamlı faktör yüküne sahip olmamasından (MTK varsayımlarının incelenmesi başlığı altında yürütülen DFA sonuçları ile birlikte incelenerek karar verilmiştir) dolayı testten çıkarılmasına karar verilmiştir. Testte bazı maddelerin c parametresinin kabul sınırları dışında olduğu ama a parametrelerinin yüksek olmasından dolayı testte yer alması kararlaştırılmıştır. 48 maddenin yeterli ayırt ediciliğe sahip olduğu

ve c parametrelerinin kabul sınırları içinde olmasından dolayı post-hoc simülasyonlarına alınmasına karar verilmiştir.

Son durumda ALES sayısal alt testin madde havuzunda 150 madde; sözel alt testin madde havuzunda ise 141 madde bulunmaktadır. Bu alt testlere ilişkin madde havuzlarının betimsel istatistikleri Tablo 21'de verilmiştir.

Tablo 23

Sayısal ve Sözel Alt Testlere Ait Madde Havuzlarının Betimsel İstatistikleri

Parametre	Madde Havuzu	Ortalama	Std. Sapma	Minimum	Maksimum
A	Sayısal	1.954	1.002	0.481	4.835
	Sözel	1.302	1.336	0.368	8.309
B	Sayısal	-1.165	0.783	-3.272	1.251
	Sözel	-0.898	1.245	-3.635	3.114
C	Sayısal	0.087	0.094	0.000	0.371
	Sözel	0.124	0.163	0.000	0.655

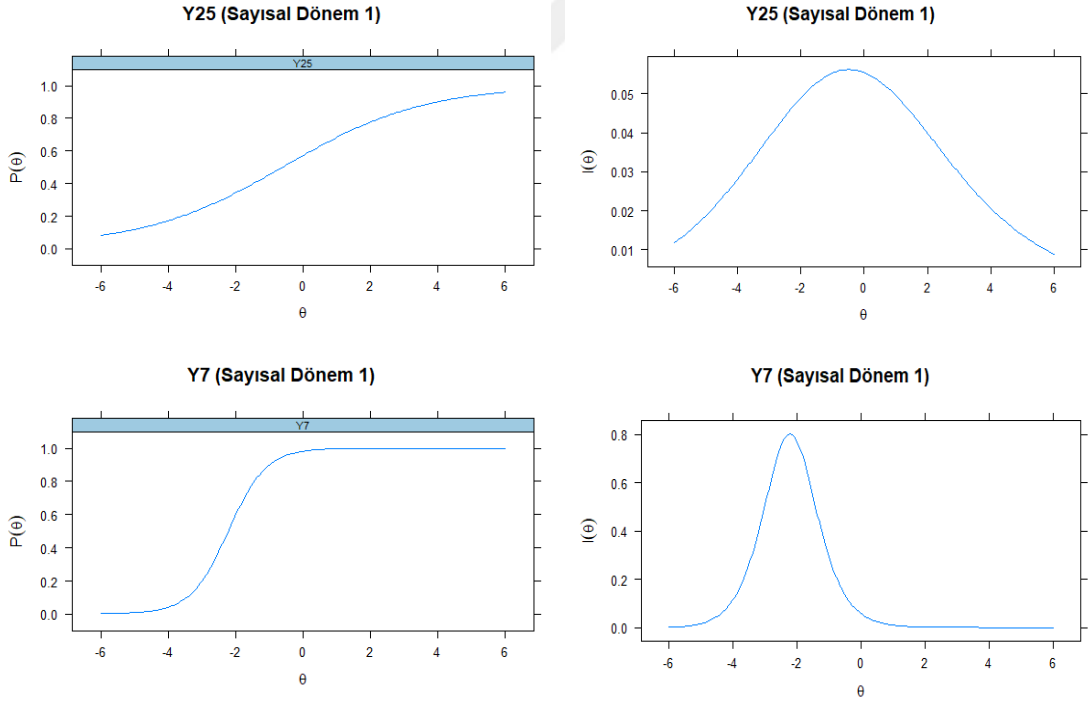
Tablo 21 incelendiğinde, sayısal alt testteki maddeler, ortalama 1.954 (ss=1.002), a parametresi ile daha yüksek ayırt ediciliğe sahiptir. Sözel alt testte ise ortalama ayırt edicilik değeri 1.302 (ss=1.336) şeklindedir. Bu durum, sözel alt testte yer alan maddelerin sayısal alt testte göre ayırt edicilik bakımından daha büyük farklılıklar olduğunu ortaya koymaktadır.

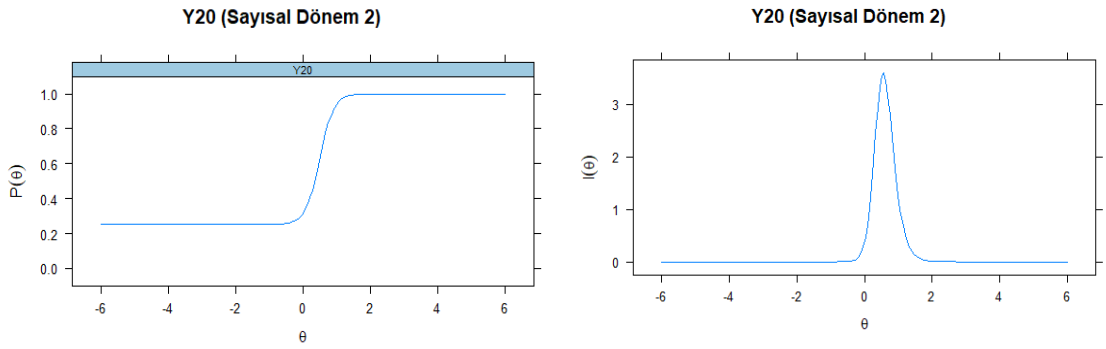
Her iki testte de ortalama b parametresi negatiftir (Sayısal: -1.165; Sözel: -0.898). Bu durum testlerin genel olarak düşük yetenek düzeyindeki bireyler için uygun olduğunu göstermektedir. Sayısal testte yer alan maddeler güçlük açısından sözel testte göre daha homojendir (Sayısal ss=0.783, Sözel ss=1.245). Bu durum sözel testte hem çok kolay hem de çok zor maddelerin bulunduğunu göstermektedir. Sözel testte maksimum b değeri 3.114 olup, bazı maddelerin yüksek yetenek düzeyindeki bireylere uygun olduğunu göstermektedir. Sayısal testteki maddelerin şans başarısı ortalaması 0.087 olup, düşük ve homojen bir dağılıma sahiptir (ss=0.094). Sözel testte ise bu değer 0.124 olup, daha yüksek ve değişkenlik gösteren bir dağılıma sahiptir (ss=0.163).

Madde sayısının yanı sıra, yeterli bir madde havuzu oluşturmak, düşük, orta ve yüksek başarı düzeyindeki katılımcı gruplarına yönelik yeterli sayıda madde içermeyi de gerektirir. PIRLS 2016 kalibre edilmiş madde havuzuna odaklanıldığında, madde güçlükleri −3.33 ile 1.14 arasında değişmektedir (Mullis vd., 2017) ve 178 maddenin 156'sı (%87'si) −2.00 ile 1.00 güçlük düzeyleri arasındadır. BOBUT uygulamalarında en

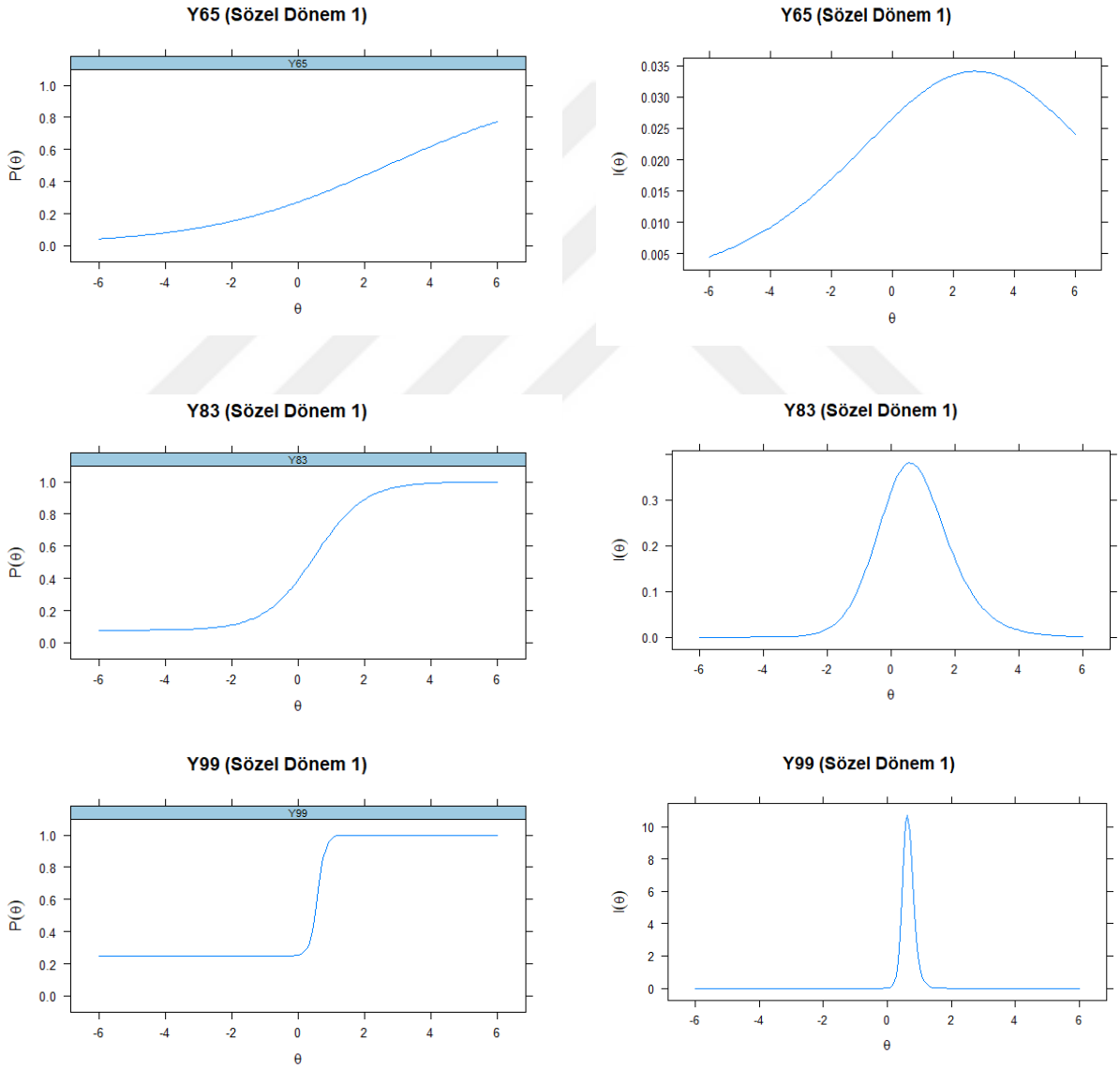
kritik sınırlamalardan biri, yetenek düzeylerinin uç noktalarında yeterli sayıda maddenin bulunmamasıdır (Valdivia Medinaceli, 2023). Bu çalışmada kullanılan madde havuzu da özellikle sayısal alt test için yetenek düzeylerinin uç noktaları için yeterli sayıda madde bulundurmamaktadır.

Madde havuzunun oluşturulmasında en önemli göstergelerden birisi de madde ve test bilgi fonksiyonlarıdır (Uttaro ve Lehman, 1999). Araştırma kapsamında ALES madde havuzundaki maddeler için Madde Karakteristik Eğrisi (MKE), Madde Bilgi Fonksiyonu (MBF) ve Test Bilgi Fonksiyonu (TBF) grafikleri üzerinden de madde havuzunun yapısı alt testlerde incelenmiştir. ALES'in sayısal ve sözel alt testte yer alan maddelerinin eğim parametreleri dikkate alınarak sayısal alt testin madde havuzundaki en düşük (Y25-dönem 1), ortanca (Y7-dönem 1) ve en yüksek (Y20-dönem 2) ayırt ediciliğe sahip maddeleri için MKE ve MBF grafikleri Şekil 10'da verilmiştir. Sözel alt testin madde havuzundaki en düşük (Y65-dönem 1), ortanca (Y83-dönem 1) ve en yüksek (Y99-dönem 2) ayırt ediciliğe sahip maddeleri için MKE ve MBF grafikleri Şekil 11'de verilmiştir.





Şekil 10. Sayısal Alt Teste Ait MKE ve MBF Örnekleri

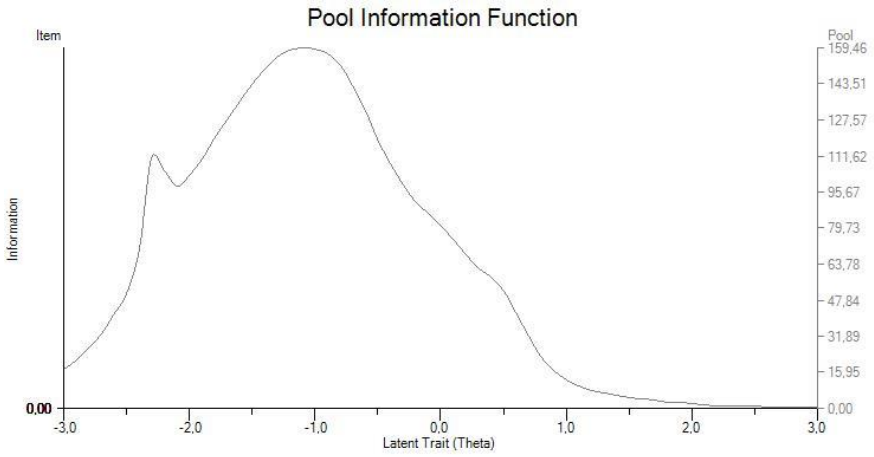


Şekil 11. Sözel Alt Teste Ait MKE ve MBF Örnekleri

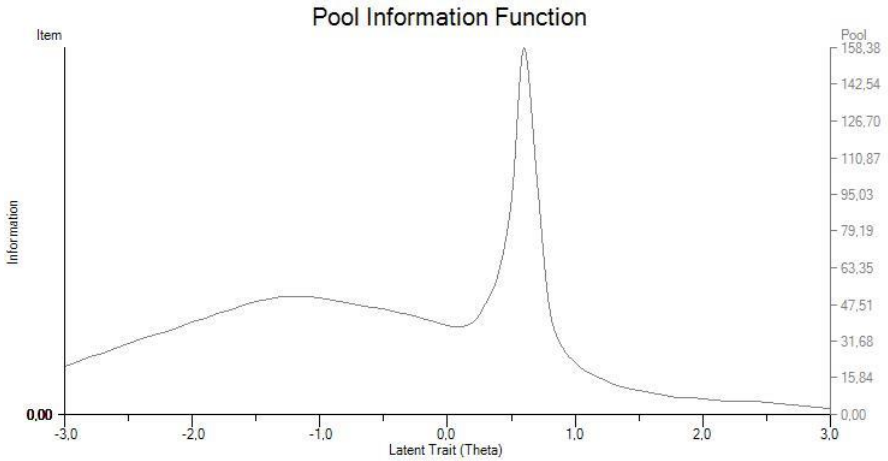
Şekil 10 ve Şekil 11'de verilen sayısal ve sözel alt testlere ait maddelerin eğimi arttıkça bilgi düzeyinin de arttığı görülmektedir. Maddelerin eğimlerinin dikleştiği

noktalarda yer alan yetenek düzeyleri için madde daha fazla bilgi sağlamıştır. Madde ayırt edicilik değeri arttıkça, bireylerin yetenek düzeyleri daha düşük hata ile kestirilebilir ve madde daha fazla bilgi sağlar (Uttaro ve Lehman, 1999). İlgili yetenek düzeyi çevresinden uzaklaştıkça maddenin sağladığı bilgi oranı azalmaktadır. Baker (2001) tarafından belirlenen kritik değerler dikkate alındığında sayısal alt test madde havuzundaki maddelerin %80'ninin (n=120) yüksek ($a \geq 1.35$), %19.3'ünün (n=29) orta düzeyde ($0.65 < a < 1.35$), %0.7'sinin (n=1), düşük ($a \leq 0.65$) ayırt ediciliğe sahip olduğu belirlenmiştir. Sözel alt test madde havuzundaki maddelerin %27.6'sının (n=39) yüksek ($a \geq 1.35$), %59.5'inin (n=84) orta düzeyde ($0.65 < a < 1.35$), %12.7'sinin (n=18), düşük ($a \leq 0.65$) ayırt ediciliğe sahip olduğu belirlenmiştir. Madde bilgi fonksiyonu, bir maddenin ölçmekte olduğu özelliğe ilişkin ne derece hassas bilgi sağladığını gösterir. Literatürde de vurgulandığı üzere (DeMars, 2010; Hambleton vd., 1991; Uttaro ve Lehman, 1999), madde ayırt ediciliği ile madde bilgi fonksiyonu arasında doğrudan bir ilişki bulunmaktadır. Ayırt edicilik değeri yüksek olan maddeler, bireyler arasındaki yetenek farklarını daha iyi ayırt ettiği için, bilgi düzeyi de daha yüksektir. Bu nedenle, bir testin toplam bilgi düzeyini artırmak amacıyla, ayırt ediciliği yüksek maddelerin seçimi büyük önem taşımaktadır.

Test bilgi fonksiyonu, madde bilgi fonksiyonlarının toplanmasıyla elde edilmektedir (Embretson ve Reise, 2000). Bu sebeple alt testlerin, yetenek düzeyine göre toplamda ne kadar bilgi sağladığını incelemek için Test Bilgi Fonksiyonları (TBF) incelenmiştir. Sayısal alt testte ilişkin TBF Şekil 12'de, Sözel alt teste ilişkin TBF ise Şekil 13'te verilmiştir.



Şekil 12. Sayısal Alt Teste Ait Test Bilgi Fonksiyonu



Şekil 13. Sözel Alt Teste Ait Test Bilgi Fonksiyonu

Şekil 12 ve Şekil 13' te verilen TBF'ler ayrı ayrı incelendiğinde alt testlerin en çok bilgi verdikleri θ düzeyinin değişmekte olduğu görülmektedir. Sayısal alt testte, en yüksek bilgi düzeyine yaklaşık θ ≈ -1.0 civarında ulaşmakta olup, bu durum testin düşük ve orta düzey yetenekli bireyleri daha hassas ölçtüğünü göstermektedir. Bilgi miktarı, yetenek düzeyi arttıkça özellikle θ > 1.0 sonrası için azalmaktadır. Testin daha çok orta ve düşük yetenek düzeylerinde daha hassas ve güvenilir kestirimler yaptığı, yüksek yetenek düzeyindeki bireyler için ek zor maddelere ihtiyaç olabileceğini göstermektedir. Sözel alt testte, testin sağladığı bilgi özellikle θ ≈ 0.75 civarında çok keskin ve yüksek bir zirveye ulaşmakta, bu da testin ortalama ve ortalama üstü yetenek düzeyindeki bireyleri en hassas şekilde ölçtüğünü göstermektedir. Ancak bilgi eğrisi bu noktadan hem sola hem sağa doğru düşmektedir. Bir başka ifadeyle çok düşük ya da çok yüksek yetenek düzeylerinde bilgi seviyesi ve dolayısıyla testin hassasiyeti azalmaktadır.

ALES Post-Hoc Simülasyonlarında Farklı Yetenek Kestirim, Test Sonlandırma ve Madde Seçim Yöntemlerine Göre Ölçme Hassasiyetinin Karşılaştırılması

Araştırmanın ikinci alt amacına doğrultusunda, madde kullanım sıklığı ve kapsam dengeleme yöntemleri tüm koşullarda sabit tutulduğunda, yetenek kestirim yöntemi olarak iki farklı yöntem (Beklenen Sonsal Dağılım [BSD], Maksimum Olabilirlik [MO]), test sonlandırma kuralı olarak beş farklı yöntem (Standart hata SH<0.30, < 0.40 ve <0.50, sabit madde sayısı MS=15, MS=20) ve madde seçim yöntemi olarak dört faklı yöntem (Maksimum Fisher bilgisi, Kullback-Leibler Bilgisi, Ağırlıklandırılmış Olabilirlik Bilgi, Verimlilik Dengeli Bilgi) kullanıldığında; ALES post-hoc simülasyonlarında uygulanan

ortalama madde sayısı ve madde havuzu kullanım oranları, ortalama standart hata değerleri, RMSE, yanlılık ve Marjinal güvenirlik değerleri, ALES post-hoc simülasyonu ve ALES asıl formdan elde edilen θ düzeyleri arasındaki ilişki incelenmiştir.

Sayısal alt test post hoc simülasyonlarından elde edilen bulgular sonlandırma kuralları göz önünde bulundurularak sekiz koşul bir arada olacak şekilde Tablo 22 - Tablo 26 arasında sunulmuştur. Tablolarda standart hata ortalamaları, uyumluluk, yanlılık ve RMSE katsayıları, marjinal güvenirlik, ortalama madde sayısı ile madde havuzunun kullanım oranları sunulmuştur.

Tablo 24

Sonlandırma Kuralı SH<0.30 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

Sonlandırma kuralı SH< 0.30								
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.303	0.300	0.283	0.369	0.264	0.264	0.254	0.261
RMSE	0.362	0.365	0.348	0.458	0.278	0.259	0.256	0.264
Yanlılık	0.047	0.050	0.051	0.079	0.025	0.019	0.029	0.014
Uyum	0.938	0.935	0.940	0.913	0.955	0.962	0.962	0.959
OMS	17.4	17.2	16.5	18.6	16.7	16.9	16.7	17.2
Havuz Kullanımı (%)	90.6	87.3	88.0	92.0	91.3	86.0	87.3	91.3
MR	0.908	0.910	0.919	0.863	0.930	0.930	0.935	0.931
Koşul	Koşul 1	Koşul 2	Koşul 3	Koşul 4	Koşul 5	Koşul 6	Koşul 7	Koşul 8

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 22'ye göre, sonlandırma kuralı SH < 0.30 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 7: BSD + AOB (0.254) yönteminde görülürken, en yüksek OSH ise Koşul 4: MO + VDB (0.369) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 7: BSD + AOB (0.256) yönteminde, en yüksek ise Koşul 4: MO + VDB (0.458) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 8: BSD + VDB (0.014) yönteminde ortaya çıkarken, en yüksek

yanlılık ise Koşul 4: MO + VDB (0.079) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu koşul Koşul 6: BSD + KLB (0.962) ve Koşul 7: BSD + AOB (0.962) olarak belirlenmiştir; en düşük uyum ise Koşul 4: MO + VDB (0.913) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı Koşul 3: MO + AOB (16.5) koşulunda gözlenirken, en yüksek ortalama madde sayısı ise Koşul 4: MO + VDB (18.6) koşulunda ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler Koşul 4: MO + VDB (%92), Koşul 8: BSD + VDB (%91.3) ve Koşul 5: BSD + MFB (%91.3) olurken, havuz kullanımı açısından en düşük oran Koşul 6: BSD + KLB (%86.0) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 7: BSD + AOB (0.935), en düşük Koşul 4: MO + VDB (0.863) koşullarından elde edilmiştir.

Tablo 25

Sonlandırma Kuralı SH<0.40 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı SH< 0.40							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.322	0.322	0.303	0.379	0.280	0.277	0.264	0.279
RMSE	0.376	0.369	0.371	0.468	0.281	0.268	0.260	0.271
Yanlılık	0.043	0.050	0.061	0.090	0.027	0.025	0.020	0.021
Uyum	0.932	0.934	0.936	0.909	0.955	0.958	0.961	0.958
OMS	14.6	14.3	14.5	17.6	14.2	13.8	13.9	14.6
Havuz Kullanımı (%)	90.0	86.0	86.0	88.0	88.0	82.0	85.3	87.3
MR	0.896	0.896	0.908	0.856	0.921	0.923	0.930	0.922
Koşullar	Koşul 9	Koşul 10	Koşul 11	Koşul 12	Koşul 13	Koşul 14	Koşul 15	Koşul 16

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 23'e göre, sonlandırma kuralı SH < 0.40 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 15: BSD + AOB (0.264) yönteminde

görülürken, en yüksek OSH ise Koşul 12: MO + VDB (0.379) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 15: BSD + AOB (0.260) yönteminde, en yüksek ise Koşul 12: MO + VDB (0.468) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 15: BSD + AOB (0.020) yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 12: MO + VDB (0.090) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu koşul Koşul 15: BSD + AOB (0.961) olarak belirlenmiştir; en düşük uyum ise Koşul 12: MO + VDB (0.909) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı Koşul 14: BSD + KLB (13.8) koşulunda gözlenirken, en yüksek ortalama madde sayısı ise Koşul 12: MO + VDB (17.6) koşulunda ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO + MFB (%90.0), MO + VDB (%88.0) ve BSD + MFB (%88.0) olurken, havuz kullanımı açısından en düşük oran BSD + KLB (%82.0) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 15: BSD + AOB (0.930), en düşük Koşul 12: MO + VDB (0.856) koşullarından elde edilmiştir.

Tablo 26

Sonlandırma Kuralı SH<0.50 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı SH< 0.50							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.345	0.340	0.313	0.374	0.297	0.295	0.280	0.298
RMSE	0.387	0.379	0.363	0.433	0.287	0.280	0.270	0.284
Yanlılık	0.058	0.062	0.050	0.066	0.026	0.016	0.017	0.018
Uyum	0.929	0.931	0.939	0.920	0.954	0.954	0.958	0.953
OMS	12.4	12.2	12.2	16.6	11.8	11.6	11.6	12.4
Havuz Kullanımı (%)	90.0	85.3	86.0	86.6	87.3	82.0	85.3	87.3
MR	0.881	0.884	0.902	0.860	0.911	0.913	0.921	0.911
Koşullar	Koşul 17	Koşul 18	Koşul 19	Koşul 20	Koşul 21	Koşul 22	Koşul 23	Koşul 24

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback -Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 24'e göre, sonlandırma kuralı SH < 0.50 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 23: BSD + AOB (0.280) yönteminde görülürken, en yüksek OSH ise Koşul 20: MO + VDB (0.374) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 23: BSD + AOB (0.270) yönteminde, en yüksek ise Koşul 20: MO + VDB (0.433) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 22: BSD + KLB (0.016) yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 20: MO + VDB (0.066) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu koşul Koşul 23: BSD + AOB (0.958) olarak belirlenmiştir; en düşük uyum ise Koşul 20: MO + VDB (0.920) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı BSD + KLB ve BSD + AOB (11.6) koşullarında gözlenirken, en yüksek ortalama madde sayısı ise MO + VDB (16.6) koşulunda ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO + MFB (%90.0), BSD + VDB (%87.3) ve BSD + MFB (%87.3) olurken, havuz kullanımı açısından en düşük oran BSD + KLB (%82.0) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik BSD + AOB (0.921), en düşük MO + VDB (0.860) koşullarından elde edilmiştir.

Tablo 27

Sonlandırma Kuralı MS=15 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı MS=15							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.361	0.361	0.376	2.036	0.260	0.260	0.254	0.307
RMSE	0.468	0.477	0.466	1.036	0.239	0.251	0.244	0.269
Yanlılık	0.108	0.114	0.101	0.492	0.033	0.034	0.033	0.048
Uyum	0.918	0.913	0.920	0.863	0.968	0.965	0.967	0.961

(Devam ediyor)

Tablo 25 (Devam)

Sonlandırma Kuralı MS=15 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı MS=15							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OMS	15	15	15	15	15	15	15	15
Havuz Kullanımı (%)	85.3	78.0	82.0	85.3	82.0	74.6	80.6	86.0
MR	0.869	0.869	0.858	-	0.932	0.932	0.935	0.905
Koşullar	Koşul 25	Koşul 26	Koşul 27	Koşul 28	Koşul 29	Koşul 30	Koşul 31	Koşul 32

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 25'e göre, sonlandırma kuralı MS=15 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 31: BSD + AOB (0.254) yönteminde görülürken, en yüksek OSH ise Koşul 28: MO + VDB (2.036) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 31: BSD + AOB (0.244) yönteminde, en yüksek ise Koşul 28: MO + VDB (1.036) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 31: BSD + AOB (0.033) ve Koşul 29: BSD + MFB (0.033) yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 28: MO + VDB (0.492) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu koşul Koşul 29: BSD + MFB (0.968) olarak belirlenmiştir; en düşük uyum ise Koşul 28: MO + VDB (0.863) yöntemlerinde hesaplanmıştır.

Madde havuzunun en etkin kullanıldığı yöntem Koşul 32: BSD + VDB (%86.0) olurken, havuz kullanımı açısından en düşük oran Koşul 30: BSD + KLB (%74.6) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 31: BSD + AOB (0.935) koşulundan, en düşük ise referans dışı olarak Koşul 28: MO + VDB (-) koşullundan elde edilmiştir.

Tablo 28

Sonlandırma Kuralı MS=20 Koşulu Altında Sayısal Alt Test Post-hoc Sonuçları

Sonlandırma kuralı MS=20								
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.296	0.296	0.295	1.544	0.225	0.228	0.226	0.239
RMSE	0.398	0.419	0.413	0.953	0.217	0.217	0.224	0.233
Yanlılık	0.082	0.084	0.087	0.419	0.027	0.027	0.033	0.029
Uyum	0.937	0.929	0.929	0.868	0.974	0.974	0.972	0.970
OMS	20	20	20	20	20	20	20	20
Havuz Kullanımı (%)	91.3	86.6	89.3	91.3	90.6	85.3	87.3	94.0
MR	0.912	0.912	0.913	-	0.949	0.948	0.948	0.942
Koşullar	Koşul 33	Koşul 34	Koşul 35	Koşul 36	Koşul 37	Koşul 38	Koşul 39	Koşul 40

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 26'ya göre, sonlandırma kuralı MS=20 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 37: BSD + MFB (0.225) yönteminde görülürken, en yüksek OSH ise Koşul 36: MO + VDB (1.544) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 37: BSD + MFB (0.217) ve Koşul 38: BSD + KLB (0.217) yöntemlerinde, en yüksek ise Koşul 36: MO + VDB (0.953) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 37: BSD + MFB (0.027) ve Koşul 38: BSD + KLB (0.027) yöntemlerinde ortaya çıkarken, en yüksek yanlılık ise Koşul 36: MO + VDB (0.419) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu koşullar Koşul 37: BSD + MFB (0.974) ve Koşul 38: BSD + KLB (0.974) olarak belirlenmiştir; en düşük uyum ise Koşul 36: MO + VDB (0.868) yöntemlerinde hesaplanmıştır.

Madde havuzunun en etkin kullanıldığı koşul Koşul 40: BSD + VDB (%94.0) olurken, havuz kullanımı açısından en düşük oran Koşul 38: BSD + KLB (%85.3) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan

MR değerleri incelendiğinde en yüksek güvenirlik Koşul 37: BSD + MFB (0.949) koşulundan, en düşük ise referans dışı olarak Koşul 36: MO + VDB (-) koşullundan elde edilmiştir.

ALES sayısal alt testinin veri seti üzerinde yapılan post-hoc simülasyonlarının bulguları birlikte yorumlandığında, madde seçim yöntemleri, yetenek kestirim yöntemleri ve test sonlandırma kuralları arasındaki etkileşimler açısından bazı dikkat çekici bulgular öne çıkmaktadır. BSD yöntemi MO yöntemine kıyasla, ortalama standart hata, RMSE, yanlılık ve marjinal güvenirlik (MR) açısından daha iyi performans göstermektedir. Özellikle sabit test uzunluklarında (MS=15, MS=20), ölçme kesinliği bakımından BSD kestirimi MO'ya göre daha iyi sonuçlar vermektedir.

VDB madde seçim yöntemi, özellikle sabit test uzunluğu ile sonlandırma koşullarında, MO yöntemiyle birlikte kullanıldığında ciddi hatalara (referans dışı) yol açmıştır. Standart hata kriterine dayalı sonlandırma uygulandığında, ortalama standart hata, RMSE ve yanlılık değerleri diğer madde seçim yöntemlerine göre yüksek çıkarken; sabit madde sayısına dayalı sonlandırmalarda (MS=15, MS=20) bu değerler belirgin şekilde artmıştır. Bunun nedeni, VDB yönteminin testin başında düşük ayırt ediciliğe sahip maddeleri, test ilerledikçe ise yüksek ayırt ediciliğe sahip maddeleri daha sık seçmesi (Han, 2012) sonucunda bireylerin yetenek düzeyi için yeterli ölçümsel bilgiye ulaşamadan testin sona ermesi olabilir. Bu durum, ortalama standart hata, RMSE ve yanlılık değerlerinin belirgin şekilde yükselmesine neden olmuştur. Referans dışı kestirim hataları MO yetenek kestirim yönteminde gözlenmekle birlikte BSD yönteminde gözlemlenmemiştir. MO yetenek kestirim yöntemi, özellikle sabit madde sayılı testlerde ve uç yetenek düzeylerinde yeterli bilgi sağlanamadığında sapmaya daha yatkındır (Embretson ve Reise, 2000; van der Linden ve Glas, 2000). Simülasyonların yürütüldüğü madde havuzunda da uç yetenek düzeyindeki bireyler için yeterince madde bulundurulmaması bu duruma neden olmuş olabilir. VDB yöntemi, testin erken aşamalarında ayırt ediciliği düşük maddelerin kullanımını da teşvik ettiği için, diğer yöntemlere kıyasla daha fazla madde kullanarak test uzunluğunu artırdığı sonucuna ulaşılmıştır. Ayrıca, madde havuzundan daha fazla sayıda madde seçtiği için havuz kullanım oranı da yüksektir. Ancak, madde havuzunu verimli kullanmasına rağmen, ölçme kesinliği açısından VDB yöntemi diğer yöntemlerin gerisinde kalmıştır.

AOB madde seçim yöntemi tüm koşullarda genellikle iyi performans sağlamakla birlikte AOB, MFB ve KLB madde seçim yöntemleri BSD yetenek kestirimi ile

kullanıldığında hem sabit test uzunluğu hem de standart hataya dayalı sonlandırma koşullarının olduğu testlerde yüksek güvenirlik ve düşük hata sağlamaktadır.

Araştırmanın ikinci alt amacına doğrultusunda sözel alt test post hoc simülasyonlarından elde edilen bulgular sonlandırma kuralları göz önünde bulundurularak sekiz koşul bir arada olacak şekilde Tablo 27 – Tablo 31 arasında sunulmuştur. Tablolarda standart hata ortalamaları, uyumluluk, yanlılık ve RMSE katsayıları, marjinal güvenirlik, ortalama madde sayısı ile madde havuzunun kullanım oranları sunulmuştur.

Tablo 29

Sonlandırma Kuralı SH<0.30 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı SH< 0.30							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.289	0.283	0.278	0.281	0.288	0.285	0.285	0.284
RMSE	0.311	0.335	0.334	0.351	0.295	0.299	0.291	0.305
Yanlılık	0.021	0.038	0.026	0.041	-0.002	0.001	0.001	0.008
Uyum	0.948	0.940	0.940	0.934	0.948	0.947	0.940	0.944
OMS	20.6	20.3	19.4	21.1	20.0	20.3	20.4	21.4
Havuz Kullanımı (%)	92.2	90.7	90.0	96.4	91.4	88.6	90.0	96.4
MR	0.916	0.919	0.922	0.921	0.917	0.918	0.918	0.919
Koşullar	Koşul 1	Koşul 2	Koşul 3	Koşul 4	Koşul 5	Koşul 6	Koşul 7	Koşul 8

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 27'ye göre, sonlandırma kuralı SH < 0.30 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 3: MO + AOB (0.278) yönteminde görülürken, en yüksek OSH ise Koşul 1: MO + MFB (0.289) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 7: BSD + AOB (0.291) yönteminde, en yüksek ise Koşul 4: MO + VDB (0.351) yönteminde ortaya çıkmıştır. Yanlılık açısından,

en düşük sapma Koşul 6: BSD + KLB (0.001) ve Koşul 7: BSD + AOB (0.001) yöntemlerinde ortaya çıkarken, en yüksek yanlılık ise Koşul 4: MO + VDB (0.041) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu madde seçim yöntemi Koşul 1: MO + MFB (0.948) ve Koşul 5: BSD + MFB (0.948) olarak belirlenmiştir; en düşük uyum ise Koşul 4: MO + VDB (0.934) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı Koşul 3: MO + AOB (19.4) madde seçim yönteminde gözlenirken, en yüksek ortalama madde sayısı ise Koşul 8: BSD + VDB (21.4) yönteminde ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO ve BSD koşulları altında Koşul 4 ve Koşul 8: VDB (%96.4) olurken, havuz kullanımı açısından en düşük oran Koşul 6: BSD + KLB (%88.6) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 3: MO + AOB (0.922), en düşük Koşul 1: MO + MFB (0.916) koşullarından elde edilmiştir.

Tablo 30

Sonlandırma Kuralı SH<0.40 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı SH< 0.40							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.362	0.350	0.341	0.350	0.360	0.357	0.354	0.358
RMSE	0.386	0.389	0.388	0.409	0.345	0.349	0.332	0.356
Yanlılık	0.016	0.028	0.003	0.021	-0.002	0.000	-0.012	-0.001
Uyum	0.918	0.919	0.919	0.910	0.929	0.936	0.934	0.923
OMS	13.9	13.9	13.5	14.9	12.9	13.2	13.1	14.1
Havuz Kullanımı (%)	87.2	85.1	85.8	90.7	85.1	82.9	85.1	91.4
MR	0.868	0.877	0.883	0.877	0.870	0.872	0.874	0.871
Koşullar	Koşul 9	Koşul 10	Koşul 11	Koşul 12	Koşul 13	Koşul 14	Koşul 15	Koşul 16

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 28'e göre, sonlandırma kuralı SH <0.40 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 11: MO + AOB (0.341) yönteminde görülürken, en yüksek OSH ise Koşul 9: MO + MFB (0.362) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 15: BSD + AOB (0.332) yönteminde, en yüksek ise Koşul 12: MO + VDB (0.409) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 14: BSD + KLB (0.000) yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 10: MO + KLB (0.028) yöntemlerinde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu madde seçim yöntemi Koşul 14: BSD + KLB (0.936) olarak belirlenmiştir; en düşük uyum ise Koşul 12: MO + VDB (0.910) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı Koşul 13: BSD + MFB (12.9) madde seçim yönteminde gözlenirken, en yüksek ortalama madde sayısı ise Koşul 12: MO + VDB (14.9) yönteminde ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO ve BSD koşulları altında Koşul 12 ve Koşul 16: VDB (%90.7 ile %91.4) olurken, havuz kullanımı açısından en düşük oran Koşul 14: BSD + KLB (%82.9) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 11: MO + AOB (0.883), en düşük Koşul 9: MO + MFB (0.868) koşullarından elde edilmiştir.

Tablo 31

Sonlandırma Kuralı SH<0.50 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

Sonlandırma kuralı SH< 0.50								
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.405	0.395	0.380	0.405	0.396	0.398	0.395	0.418
RMSE	0.431	0.421	0.406	0.436	0.378	0.365	0.361	0.391
Yanlılık	0.038	0.026	0.001	-0.001	-0.001	0.008	0.004	0.006
Uyum	0.905	0.908	0.913	0.902	0.913	0.920	0.901	0.907

(Devam ediyor)

Tablo 29 (Devam)

Sonlandırma Kuralı SH<0.50 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

Sonlandırma kuralı SH< 0.50								
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OMS	11.5	11.4	11.1	12.0	10.7	10.8	10.7	11.4
Havuz Kullanımı (%)	85.1	82.2	81.5	90.7	83.6	78.7	75.8	89.3
MR	0.835	0.843	0.855	0.835	0.843	0.841	0.843	0.825
Koşullar	Koşul 17	Koşul 18	Koşul 19	Koşul 20	Koşul 21	Koşul 22	Koşul 23	Koşul 24

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 29'a göre, sonlandırma kuralı SH <0.50 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 19: MO + AOB (0.380) yönteminde görülürken, en yüksek OSH ise Koşul 24: BSD + VDB (0.418) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 23: BSD + AOB (0.361) yönteminde, en yüksek ise Koşul 20: MO + VDB (0.436) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 19: MO + AOB (0.001), Koşul 20: MO + VDB (-0.001) ve Koşul 21: BSD + MFB (-0.001) yöntemlerinde ortaya çıkarken, en yüksek yanlılık ise Koşul 17: MO + MFB (0.038) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu madde seçim yöntemi Koşul 22: BSD + KLB (0.920) olarak belirlenmiştir; en düşük uyum ise Koşul 23: BSD + AOB (0.901) yöntemlerinde hesaplanmıştır.

En düşük ortalama madde sayısı Koşul 21: BSD + MFB (10.7) ve Koşul 23: BSD + AOB (10.7) madde seçim yöntemlerinde gözlenirken, en yüksek ortalama madde sayısı ise Koşul 20: MO + VDB (12.0) yönteminde ortaya çıkmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO ve BSD koşulları altında Koşul 20 ve Koşul 24: VDB (%90.7 ile %89.3) olurken, havuz kullanımı açısından en düşük oran Koşul 23: BSD + AOB (%75.8) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği

için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 19: MO + AOB (0.855), en düşük Koşul 24: BSD + VDB (0.825) koşullarından elde edilmiştir.

Tablo 32

Sonlandırma Kuralı MS=15 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı MS=15							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.353	0.342	0.333	0.376	0.333	0.339	0.337	0.357
RMSE	0.409	0.405	0.373	0.432	0.315	0.315	0.319	0.323
Yanlılık	0.053	0.036	0.027	0.047	0.001	0.014	0.012	0.011
Uyum	0.922	0.923	0.932	0.914	0.941	0.941	0.943	0.938
OMS	15	15	15	15	15	15	15	15
Havuz Kullanımı (%)	82.2	75.8	76.5	93.6	80.1	75.1	77.3	91.4
MR	0.875	0.883	0.889	0.859	0.889	0.885	0.886	0.873
Koşullar	Koşul 25	Koşul 26	Koşul 27	Koşul 28	Koşul 29	Koşul 30	Koşul 31	Koşul 32

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 30'a göre, sonlandırma kuralı MS=15 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 27: MO + AOB (0.333) ve Koşul 29: BSD + MFB (0.333) yöntemlerinde görülürken, en yüksek OSH ise Koşul 28: MO+ VDB (0.376) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 29: BSD + MFB ve Koşul 30: BSD + KLB (0.315) yöntemlerinde, en yüksek ise Koşul 28: MO + VDB (0.432) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 29: BSD + MFB (0.001), yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 25: MO + MFB (0.053) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu madde seçim yöntemi Koşul 31: BSD + AOB (0.943) olarak belirlenmiştir; en düşük uyum ise Koşul 28: MO + VDB (0.914) yöntemlerinde hesaplanmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO ve BSD koşulları altında Koşul 28 ve Koşul 32: VDB (%93.6 ile %91.4) olurken,

havuz kullanımı açısından en düşük oran Koşul 30: BSD + KLB (%75.1) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 29: BSD + MFB (0.889) ve Koşul 27: MO + AOB (0.889), en düşük Koşul 28: MO + VDB (0.859) koşullarından elde edilmiştir.

Tablo 33

Sonlandırma Kuralı MS=20 Koşulu Altında Sözel Alt Test Post-hoc Sonuçları

	Sonlandırma kuralı MS=20							
Yetenek kestirim yöntemi	MO				BSD			
Madde Seçim Yöntemi	MFB	KLB	AOB	VDB	MFB	KLB	AOB	VDB
OSH	0.295	0.293	0.287	0.309	0.286	0.291	0.290	0.304
RMSE	0.334	0.331	0.327	0.360	0.273	0.277	0.277	0.283
Yanlılık	0.029	0.028	0.034	0.029	-0.001	0.016	0.005	0.006
Uyum	0.943	0.945	0.946	0.936	0.959	0.957	0.955	0.953
OMS	20	20	20	20	20	20	20	20
Havuz Kullanımı (%)	86.5	80.8	83.6	98.5	85.1	80.1	84.3	95.7
MR	0.913	0.914	0.918	0.905	0.918	0.915	0.916	0.908
Koşullar	Koşul 33	Koşul 34	Koşul 35	Koşul 36	Koşul 37	Koşul 38	Koşul 39	Koşul 40

MO: Maksimum Olabilirlik, BSD: Beklenen Sonsal Dağılım, MFB: Maksimum Fisher Bilgisi, KLB: Kullback-Leibler Bilgisi, AOB: Ağırlıklandırılmış Olabilirlik Bilgi Kriteri, VDB: Verimlilik Dengeli Bilgi, OSH: Ortalama Standart Hata, OMS: Ortalama Madde Sayısı, MR: Marjinal Güvenirlik

Tablo 31'e göre, sonlandırma kuralı MS=20 olarak belirlendiğinde, en düşük ortalama standart hata (OSH) değeri Koşul 37: BSD + MFB (0.286) yönteminde görülürken, en yüksek OSH ise Koşul 36: MO+ VDB (0.309) yönteminde ölçülmüştür. RMSE değerlerine bakıldığında, en düşük Koşul 37: BSD + MFB (0.273) yönteminde, en yüksek ise Koşul 36: MO + VDB (0.360) yönteminde ortaya çıkmıştır. Yanlılık açısından, en düşük sapma Koşul 37: BSD + MFB (-0.001), yönteminde ortaya çıkarken, en yüksek yanlılık ise Koşul 35: MO + AOB (0.034) yönteminde gözlemlenmiştir. Yetenek kestirimi ile gerçek yetenek değerleri arasındaki uyumun en yüksek olduğu madde seçim yöntemi Koşul 37: BSD + MFB (0.959) olarak belirlenmiştir; en düşük

uyum ise Koşul 36: MO + VDB (0.936) yöntemlerinde hesaplanmıştır. Madde havuzunun en etkin kullanıldığı yöntemler MO ve BSD koşulları altında Koşul 36 ve Koşul 40: VDB (%98.5 ile %95.7) olurken, havuz kullanımı açısından en düşük oran Koşul 38: BSD + KLB (%80.1) olarak kaydedilmiştir. Son olarak, bireylerin yetenek kestirimlerinin güvenirliği için hesaplanan MR değerleri incelendiğinde en yüksek güvenirlik Koşul 37: BSD + MFB (0.918) ve Koşul 35: MO + AOB (0.918), en düşük Koşul 36: MO + VDB (0.905) koşullarından elde edilmiştir.

ALES sözel alt testinin post-hoc simülasyon sonuçları için tüm koşullar değerlendirildiğinde dikkat çeken bazı bulgular elde edilmiştir. Standart hata (SH) sonlandırma kuralı altında, MFB yöntemi, BSD yetenek kestirimi ile; AOB, KLB ve VDB yöntemleri ise MO yöntemi ile uyumlu çalışmıştır. Bu uyum, ortalama standart hatanın (OSH) daha düşük çıkması açısından belirgindir. Sabit test uzunluğu sonlandırma kuralında ise tüm madde seçim yöntemlerinin genel olarak BSD ile daha uyumlu çalıştığı gözlemlenmiştir. Bu uyum, özellikle ortalama standart hata (OSH), RMSE, yanlılık ve uyum değerlerinin daha düşük çıkması açısından belirgindir. RMSE ve yanlılık (bias) değerleri incelendiğinde, tüm sonlandırma kuralları altında BSD yönteminin, MO yöntemine kıyasla genellikle daha düşük değerler ürettiği görülmüştür. MO yöntemi, Bayesyen kestirim yöntemlerine göre daha düşük yanlılık üretse de standart hata ve RMSE açısından daha yüksek değerlere sahip olduğu literatürde de vurgulanmıştır (Bock ve Mislevy, 1982; Crichton, 1981; Wang ve Vispoel, 1998; Wang vd., 1999; Özdemir, 2016). Buna karşın BSD yetenek kestirim yönteminin, MO yöntemine kıyasla daha düşük RMSE ve yanlılık değerleri ürettiği, kararlı kestirimler yaptığı literatürde yer alan bazı araştırmalarla da örtüşmektedir (Balta ve Uçar, 2022; Seo ve Choi, 2018; Yi vd., 2000; Wang ve Vispoel, 1998). Bu çalışmada kullanılan simülasyon koşullarının literatürdeki çalışmalardan farklılık göstermesi ve analizlerin gerçek verilere dayalı olarak gerçekleştirilmesi, MO yönteminin BSD yöntemine kıyasla daha yüksek yanlılık üretmesine neden olmuş olabilir.

Araştırma bulgularına göre, sabit uzunluklu (MS=20) sonlandırma kuralları, standart hataya (SH) dayalı sonlandırma kurallarına kıyasla ölçme kesinliği açısından daha iyi sonuçlar vermiştir. Özellikle madde havuzunun geniş bir güçlük yelpazesine sahip olmadığı durumlarda, SH'ye dayalı sonlandırma kuralları verimli bir test süreci sağlayamayabilir. Buna karşın, sabit uzunluklu testlerin daha küçük yanlılık ve standart hata değerleriyle sonuçlanabileceği literatürde de belirtilmektedir (Yi vd., 2000). Ayrıca, BOBUT uygulamalarında madde havuzundaki madde sayısının sınırlı olduğu durumlarda

sabit test uzunluklu durdurma kurallarının tercih edilmesi gerektiği vurgulanmaktadır (Babcock ve Weiss, 2009).

Bu çalışmada, VDB yöntemi tüm koşullarda en yüksek ortalama madde sayısı ve madde havuzu kullanım oranına sahip olmasıyla dikkat çekmiştir. Benzer şekilde, Lee ve Han (2024) çalışmalarında, VDB yönteminin düşük ayırt ediciliğe sahip maddeleri de etkin biçimde kullanmasından dolayı madde havuzunu en yoğun şekilde kullanan yöntem olduğunu ifade etmişlerdir. Ayrıca, bu çalışmanın bulgularıyla paralel olarak, MFB ve KLB yöntemlerinin VDB'ye kıyasla daha kısa testler ürettiği sonucuna ulaşmışlardır.

SH < 0.30 düzeyinde marjinal güvenirlik katsayısı 0.90'ın üzerinde, SH < 0.40 ve SH < 0.50 düzeylerinde ise 0.80'in üzerinde bulunmuştur. Sabit uzunluklu test sonlandırma kuralında ise, test uzunluğu arttıkça (MS=20 olunca) marjinal güvenirliğin de arttığı görülmektedir. Öte yandan, standart hata eşiği daha yüksek olan bir sonlandırma kuralı kullanıldığında, BOBUT'un sağladığı güvenirliğin belirgin şekilde azaldığı gözlemlenmiştir. Bu bulgu, Smits vd. (2011) çalışmalarıyla da örtüşmektedir. Söz konusu çalışmada daha yüksek SH eşiklerinin marjinal güvenirliği düşürdüğü rapor edilmiştir.

ALES Post-Hoc Simülasyonları Sonucunda Alt Testler İçin Ölçme Hassasiyeti Açısından En Uygun BOBUT Koşulları

Araştırmanın üçüncü alt amacı doğrultusunda, ALES post-hoc simülasyonları sonucunda alt testler için ölçme hassasiyeti açısından en uygun BOBUT koşullarına ilişkin ALES sayısal alt testine ait bulgular Tablo 32'de sunulmuştur.

Tablo 34

ALES Sayısal Alt Test Post-Hoc Değerlendirme Sonuçlarına Göre Değerlendirme Kriterlerinin En Düşük ve En Yüksek Olduğu Koşullar

Değerlendirme Kriterleri	En yüksek		En düşük	
	Değer	Yöntem	Değer	Yöntem
OSH	2.036	MO + VDB MS=15	0.225	BSD + MFB MS=20
RMSE	1.036	MO + VDB MS=15	0.217	BSD + MFB MS=20 BSD + KLB MS=20
Yanlılık	0.492	MO + VDB MS=15	0.014	BSD + VDB SH<0.30

(Devam ediyor)

Tablo 32 (Devam)

ALES Sayısal Alt Test Post-Hoc Değerlendirme Sonuçlarına Göre Değerlendirme Kriterlerinin En Düşük ve En Yüksek Olduğu Koşullar

Değerlendirme Kriterleri	En yüksek		En düşük	
	Değer	Yöntem	Değer	Yöntem
Uyum	0.974	BSD + MFB MS=20 BSD + KLB MS=20	0.863	MO + VDB MS=15
Havuz Kullanımı	94.0	BSD + VDB MS=20	74.6	BSD + KLB MS=15
MR	0.949	BSD + MFB MS=20	Referans Dışı	MO + VDB MS=15

Tablo 32 incelendiğinde farklı sonlandırma koşulları, madde seçim yöntemleri ve yetenek kestirim yöntemleri altında en düşük OSH 0.225 değeri ile BSD + MFB; MS=20 koşulunda, en yüksek OSH ise 2.036 ile MO + VDB; MS=15 koşulunda bulunmuştur. Yetenek kestirimlerinin doğruluğu için hesaplanan RMSE değerleri incelendiğinde en düşük değer 0.217 ile BSD + MFB; MS=20 ve BSD + KLB; MS=20 koşullarında, en yüksek 1.036 ile MO + VDB; MS=15 koşulunda bulunmuştur. Bireylerin kestirilen yetenek düzeylerinin gerçek yeteneklerinden ne kadar saptığını gösteren yanlılık değerleri incelendiğinde en düşük sapma değeri 0.014 ile BSD + VDB; SH<0.30 koşulunda, en yüksek sapma ise 0.492 ile MO + VDB; MS=15 koşulunda görülmüştür. Bireylerin gerçek yetenek düzeyi ile kestirilen yetenek düzeylerinin arasındaki doğruluğu ölçen uyum değerleri incelendiğinde, en yüksek uyum değerleri 0.974 ile BSD + MFB; MS=20 ve BSD + KLB; MS=20 koşullarında, en düşük uyum değeri ise 0.863 ile MO + VDB; MS=15 koşulunda elde edilmiştir. Havuz kullanım oranları incelendiğinde en yüksek kullanım oranları %94 ile BSD + VDB; MS=20 koşulu altında en düşük ise 74.6 ile BSD + KLB; MS=15 koşulunda gözlemlenmiştir. Marjinal güvenirlik değerleri incelendiğinde en yüksek güvenirlik değeri 0.949 ile BSD + MFB; MS=20 koşulunda en düşük güvenirlik değeri ise referans dışı olarak MO + VDB; MS=15 koşulundan elde edilmiştir.

Yapılan analizler sonucunda, BOBUT uygulamalarında kullanılan farklı madde seçim ve yetenek kestirim yöntemlerinin performansları çeşitli kriterler açısından karşılaştırılmıştır. Tüm koşullar göz önünde bulundurulduğunda, BSD + MFB; MS=20 koşullunun öne çıktığı görülmüştür. BSD + MFB; MS=20 koşulunda; ortalama standart hata (OHS) 0.225, (RMSE) 0.217, yanlılık değeri 0.027, gerçek yetenek ile kestirilen

yetenek arasındaki uyum 0.974, ortalama madde sayısı 20, havuz kullanım oranı %90.6 ve marjinal güvenirlik (MR) katsayısı 0.949 olarak hesaplanmıştır. BSD + MFB; MS=20 koşulu; daha düşük OSH ve RMSE değerlerine sahip olması, yanlılık değerinin de diğer koşullara nazaran düşük olması, daha yüksek uyum ve güvenirlik katsayısı göstermesi nedeniyle ölçme doğruluğu ve tarafsızlık açısından daha üstün bir performans sergilemiştir. Testin 20 madde ile sonlandırılması, kâğıt kalem testlerine kıyasla test uzunluğunu yaklaşık %60 oranında azaltmaktadır. Bu bağlamda, daha doğru ve güvenilir yetenek kestirimi yapılmasını sağlayan BSD + MFB; MS=20 koşulu, ALES Matematik alt testinin canlı BOBUT uygulaması için en uygun seçenek olarak değerlendirilmiştir.

Araştırmanın üçüncü alt amacı doğrultusunda ALES sözel alt test için bulgular Tablo 33'te sunulmuştur.

Tablo 35

ALES Sözel Alt Test Post-Hoc Değerlendirme Sonuçlarına Göre Değerlendirme Kriterlerinin En Düşük ve En Yüksek Olduğu Koşullar

Değerlendirme Kriterleri	En yüksek		En düşük	
	Değer	Yöntem	Değer	Yöntem
OSH	0.418	BSD + VDB SH< 0.50	0.278	MO + AOB SH<0.30
RMSE	0.436	MO + VDB SH< 0.50	0.273	BSD + MFB MS=20
Yanlılık	0.053	MO + MFB MS=15	0.000	BSD + KLB SH<0.40
Uyum	0.959	BSD + MFB MS=20	0.901	BSD + AOB SH<0.50
Havuz Kullanımı	%96.4	MO + VDB BSD + VDB SH< 0.50	%75.1	BSD + KLB MS=15
MR	0.922	MO + AOB SH<0.30	0.825	BSD + VDB SH< 0.50

Tablo 33 incelendiğinde farklı sonlandırma koşulları, madde seçim yöntemleri ve yetenek kestirim yöntemleri altında en düşük OSH 0.278 değeri ile MO + AOB; SH<0.30 koşulunda, en yüksek OSH ise 0.418 ile BSD + VDB; SH< 0.50 koşulunda bulunmuştur. Yetenek kestirimlerinin doğruluğu için hesaplanan RMSE değerleri incelendiğinde en düşük değer 0.273 ile BSD + MFB; MS=20 koşulunda, en yüksek 0.436 ile MO + VDB; SH< 0.50 koşulunda bulunmuştur. Bireylerin kestirilen yetenek düzeylerinin gerçek yeteneklerinden ne kadar saptığını gösteren yanlılık değerleri incelendiğinde en düşük

sapma değeri 0.000 ile BSD + KLB; SH<0.40 koşulunda, en yüksek sapma ise 0.053 ile MO + MFB; MS=15 koşulunda görülmüştür. Bireylerin gerçek yetenek düzeyi ile kestirilen yetenek düzeylerinin arasındaki doğruluğu ölçen uyum değerleri incelendiğinde, en yüksek uyum değerleri 0.959 ile BSD + MFB; MS=20 koşulunda, en düşük uyum değerleri ise 0.901 ile BSD + AOB; SH<0.50 koşulunda elde edilmiştir. Havuz kullanım oranları incelendiğinde en yüksek kullanım oranları %96.4 ile MO + VDB ve BSD + VDB; SH< 0.50 koşulları altında en düşük ise %75.1 ile BSD + KLB; MS=15 koşulunda gözlemlenmiştir. Marjinal güvenirlik değerleri incelendiğinde en yüksek güvenirlik değeri 0.922 ile MO + AOB; SH<0.30 koşulunda en düşük güvenirlik değeri ise 0.825 değeri ile BSD + VDB; SH< 0.50 koşulundan elde edilmiştir.

Yapılan analizler sonucunda, BOBUT uygulamalarında kullanılan farklı madde seçim ve yetenek kestirim yöntemlerinin performansları çeşitli kriterler açısından karşılaştırılmıştır. Tüm koşullar göz önünde bulundurulduğunda, BSD + MFB; MS=20 ve MO + AOB; SH<0.30 koşullarının öne çıktığı görülmüştür. Bu nedenle, söz konusu iki koşul detaylı şekilde karşılaştırılmıştır. BSD + MFB; MS=20 koşulunda; ortalama standart hata (OHS) 0.286, (RMSE) 0.273, yanlılık değeri -0.001, gerçek yetenek ile kestirilen yetenek arasındaki uyum 0.959, ortalama madde sayısı 20, havuz kullanım oranı %85.1 ve marjinal güvenirlik (MR) katsayısı 0.918 olarak hesaplanmıştır. MO + AOB; SH<0.30 koşulunda ise; OHS 0.278, RMSE 0.334, yanlılık 0.026, uyum 0.940, ortalama madde sayısı 19.4, havuz kullanım oranı %90.0 ve MR katsayısı 0.922 olarak bulunmuştur. Her iki koşul benzer performans sergilemekle birlikte, BSD + MFB; MS=20 koşulu; daha düşük RMSE ve yanlılık değerlerine sahip olması, daha yüksek uyum katsayısı göstermesi nedeniyle ölçme doğruluğu ve tarafsızlık açısından daha üstün bir performans sergilemiştir. Ayrıca her iki koşulda da ortalama madde sayısının 20 civarında olması, kâğıt kalem testlerine kıyasla test uzunluğunu yaklaşık %60 oranında azaltmaktadır. Bu bağlamda, daha doğru ve güvenilir yetenek kestirimi yapılmasını sağlayan BSD + MFB; MS=20 koşulu, ALES Türkçe alt testinin canlı BOBUT uygulaması için en uygun seçenek olarak değerlendirilmiştir.

Ayrıca, alt testler için en uygun koşul olan BSD + MFB; MS=20 altında, her bir alt testte sınav esnasında birey düzeyinde kapsam dengeleme koşulunun sağlanıp sağlanmadığı incelenmiştir. Sayısal alt testte, incelenen bireye "Sayılarla İşlem" alanından 6, "Sayısal Mantık ve Problem Çözme" alanından 11 ve "Geometri" alanından 3 madde yöneltilmiştir. Alt testin içerik alanlarına göre belirlenen hedef oranlar; Sayılarla İşlem için %30, Sayısal Mantık ve Problem Çözme için %56, Geometri için ise %14'tür.

Bireye uygulanan maddelerin içerik oranları ise sırasıyla %30, %55 ve %15 olarak gerçekleşmiştir. Bu bulgular, bireye yöneltilen maddelerin testin belirlenen içerik oranlarına büyük ölçüde uygun olarak seçildiğini ve içerik dengesinin birey düzeyinde başarıyla sağlandığını göstermektedir. Sözel alt testte, incelenen bireye "Dil Bilgisi ve Anlam Bütünlüğü" alanından 8, "Paragraf" alanından 9 ve "Sözel mantık" alanından 3 madde yöneltilmiştir. Alt testin içerik alanlarına göre belirlenen hedef oranlar; Dil Bilgisi ve Anlam Bütünlüğü için %36, Paragraf için %47, Sözel mantık için ise %17'dir. Bireye uygulanan maddelerin içerik oranları ise sırasıyla %40, %45 ve %15 olarak gerçekleşmiştir. Bu bulgular, bireye yöneltilen maddelerin testin belirlenen içerik oranlarına büyük ölçüde uygun olarak seçildiğini ve içerik dengesinin birey düzeyinde başarıyla sağlandığını göstermektedir.

ALES Canlı BOBUT Uygulaması

Araştırmanın dördüncü alt amacı kapsamında ALES post-hoc simülasyon sonuçlarına göre belirlenen en uygun yetenek kestirim yöntemi, test sonlandırma kuralı ve madde seçim yöntemi kullanılarak oluşturulan ALES canlı BOBUT uygulaması ve ALES kağıt-kalem testi uygulanmıştır. Bu kapsamda yer alan alt amaçlar ve bulgular aşağıda sırasıyla yer almaktadır.

Katılımcıların ALES Kağıt-Kalem Formundan ve ALES Canlı BOBUT Uygulamasından Kestirilen Yetenek Düzeyleri Arasındaki İlişkinin İncelenmesi

Araştırmanın dördüncü alt amacının (a) maddesi kapsamında, katılımcıların ALES kağıt-kalem formundan kestirilen ve ALES canlı BOBUT uygulamasından elde ettikleri yetenek düzeylerinin arasındaki ilişkiyi belirlemek için Pearson Momentler Çarpımı korelasyon katsayıları hesaplanmıştır. Sonuçlar sayısal ve sözel alt testler için ayrı ayrı incelenmiş olup Tablo 34'te verilmiştir.

Tablo 36

Katılımcıların ALES Alt Testlerinin Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri Yetenek Düzeyleri Arasındaki İlişki

Alt Test	Uygulama Türü	r	p
Sayısal	Canlı BOBUT KK formu	0.75	0.00
Sözel	Canlı BOBUT KK formu	0.59	0.00

Tablo 34 incelendiğinde, ALES alt testlerinde KK formu ile canlı BOBUT uygulaması arasında anlamlı ilişkiler olduğu görülmektedir. Sayısal alt testte katılımcıların KK ve BOBUT yetenek düzeyleri arasındaki korelasyon r = .75 olarak bulunmuş ve bu ilişki istatistiksel olarak anlamlı ($p < 0.05$) çıkmıştır. Bu durum, sayısal alt testte iki uygulama türü arasında yüksek düzeyde tutarlılık olduğunu ve BOBUT yetenek düzeyleri kağıt-kalem yetenek düzeyleriyle oldukça uyumlu olduğunu göstermektedir. Sözel alt testte ise korelasyon r = .59 olarak hesaplanmış ve bu ilişki de anlamlıdır ($p < 0.05$). Sözel alt test için bu orta düzeydeki ilişki, yetenek düzeyleri arasında pozitif bir bağlantı olduğunu göstermekte, ancak sayısal alt teste göre biraz daha düşük bir tutarlılık sergilemektedir. Canlı BOBUT uygulamasının kâğıt-kalem testlerinin yerine kullanılabilirliğini inceleyen önceki çalışmalarda (Chew, 1997; Hogan, 2014; Kalender, 2011; Şenel, 2017; Žitný vd., 2012) iki uygulamadan elde edilen yetenek parametreleri arasındaki korelasyonlar ise 0.53 ile 0.95 arasında değiştiği belirtilmektedir. Genel olarak, her iki alt testte de KK ve canlı BOBUT uygulamalarından elde edilen yetenek düzeyleri arasında anlamlı ve pozitif ilişkiler bulunması, BOBUT uygulamasının KK uygulamasına benzer şekilde yetenek kestirimleri sağladığını göstermektedir.

Katılımcıların ALES Kağıt-Kalem Formundan ve ALES Canlı BOBUT Uygulamasından Kestirilen Yetenek Parametreleri Arasındaki Manidarlığın İncelenmesi

Araştırmanın dördüncü alt amacının (b) maddesi kapsamında, katılımcıların ALES kağıt-kalem ve ALES canlı BOBUT uygulamasından elde ettikleri yetenek parametrelerinin ortalamaları arasındaki farkın anlamlılığını belirlemek üzere sayısal ve sözel alt testler ayrı ayrı incelenmiştir. Tablo 35'te sayısal alt test, Tablo 36'da sözel alt

test için kağıt-kalem ve BOBUT uygulamalarının yetenek kestirimleri ile standart hata değerleri verilmiştir. Tablo 37'de sayısal ve Tablo 38'de sözel alt test için kestirilen yetenek parametrelerinin ortalamaları arasındaki farkın anlamlılığının incelendiği t-testi sonuçları sunulmuştur.

Tablo 37

Katılımcıların ALES Sayısal Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri θ Düzeyleri ve Standart Hataları (SH)

Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)
1	0.81 (0.23)	1.02 (0.33)	35	0.26 (0.40)	-0.13 (0.20)	69	-1.34 (0.18)	-1.94 (0.15)
2	0.19 (0.18)	0.28 (0.20)	36	0.50 (0.32)	0.26 (0.25)	70	-1.29 (0.11)	0.14 (0.16)
3	0.14 (0.36)	-0.59 (0.14)	37	0.28 (0.16)	0.01 (0.16)	71	-1.66 (0.10)	-0.85 (0.15)
4	0.28 (0.25)	-0.15 (0.15)	38	-0.24 (0.29)	-0.47 (0.15)	72	-1.98 (0.40)	-2.18 (0.16)
5	0.29 (0.25)	0.19 (0.17)	39	-0.17 (0.23)	-0.34 (0.15)	73	0.80 (0.31)	0.53 (0.20)
6	0.17 (0.37)	-0.63 (0.15)	40	-0.05 (0.15)	-0.15 (0.19)	74	0.29 (0.30)	0.04 (0.17)
7	-0.43 (0.08)	0.03 (0.23)	41	-0.09 (0.12)	-0.70 (0.13)	75	0.15 (0.20)	-0.10 (0.16)
8	-0.21 (0.20)	-0.47 (0.24)	42	-0.71 (0.13)	-1.15 (0.16)	76	0.76 (0.31)	0.31 (0.16)
9	-0.78 (0.22)	0.25 (0.15)	43	-0.23 (0.28)	-0.45 (0.14)	77	0.74 (0.37)	0.84 (0.27)
10	0.74 (0.25)	0.25 (0.21)	44	-0.28 (0.22)	-0.67 (0.14)	78	1.37 (0.52)	1.27 (0.43)
11	0.22 (0.36)	-0.26 (0.20)	45	-2.05 (0.48)	-2.23 (0.19)	79	-0.13 (0.12)	-0.04 (0.15)
12	0.39 (0.51)	-0.34 (0.16)	46	-2.13 (0.36)	-2.10 (0.16)	80	0.10 (0.19)	0.10 (0.21)
13	0.47 (0.17)	-0.28 (0.30)	47	-2.35 (0.41)	-0.96 (0.13)	81	0.34 (0.16)	1.23 (0.41)
14	-0.12 (0.14)	0.02 (0.13)	48	-2.13 (0.39)	-2.92 (0.36)	82	0.21 (0.18)	0.27 (0.16)
15	-0.47 (0.50)	-0.77 (0.14)	49	-0.56 (0.10)	-0.89 (0.15)	83	0.01 (0.17)	1.00 (0.31)
16	0.44 (0.41)	-0.57 (0.16)	50	-1.15 (0.20)	-1.91 (0.21)	84	0.22 (0.23)	0.12 (0.18)
17	-0.50 (0.22)	-0.61 (0.14)	51	-0.28 (0.12)	-0.40 (0.14)	85	-0.12 (0.12)	0.24 (0.21)
18	-1.13 (0.28)	-0.81 (0.21)	52	-1.48 (0.18)	-1.18 (0.15)	86	-0.09 (0.20)	-0.22 (0.17)

(Devam ediyor)

Tablo 35 (Devam)

Katılımcıların ALES Sayısal Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri θ Düzeyleri ve Standart Hataları (SH)

Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)
19	-0.88 (0.38)	-1.46 (0.14)	53	-0.70 (0.18)	-0.27 (0.15)	87	0.85 (0.34)	0.32 (0.20)
20	-0.11 (0.20)	-0.18 (0.23)	54	0.68 (0.45)	-0.56 (0.15)	88	-0.10 (0.12)	0.06 (0.15)
21	-1.02 (0.40)	-1.02 (0.13)	55	-0.53 (0.10)	-0.48 (0.16)	89	0.56 (0.17)	0.47 (0.18)
22	0.26 (0.29)	-0.79 (0.14)	56	-1.39 (0.17)	-1.11 (0.15)	90	0.31 (0.13)	0.42 (0.20)
23	-3.07 (0.53)	-1.09 (0.13)	57	-0.21 (0.15)	-0.37 (0.16)	91	0.61 (0.38)	0.64 (0.21)
24	-0.43 (0.46)	-1.03 (0.13)	58	-1.33 (0.15)	-0.49 (0.14)	92	-0.40 (0.00)	0.31 (0.23)
25	-1.78 (0.64)	-1.91 (0.15)	59	0.47 (0.34)	0.12 (0.17)	93	1.47 (0.56)	0.46 (0.25)
26	-2.41 (0.28)	-0.89 (0.13)	60	0.46 (0.22)	0.03 (0.16)	94	0.16 (0.18)	0.39 (0.19)
27	-1.33 (0.65)	-1.80 (0.18)	61	-2.09 (0.48)	-1.72 (0.28)	95	-1.64 (0.62)	-1.62 (0.27)
28	0.34 (0.53)	-0.48 (0.19)	62	-2.50 (0.51)	-1.84 (0.15)	96	-0.69 (0.50)	-1.10 (0.16)
29	-1.69 (0.20)	-0.75 (0.14)	63	-0.48 (0.18)	-1.11 (0.15)	97	-0.51 (0.38)	-0.59 (0.14)
30	-1.02 (0.31)	-0.29 (0.16)	64	-0.89 (0.11)	-1.36 (0.20)	98	-1.34 (0.19)	-1.24 (0.14)
31	-0.63 (0.09)	-0.72 (0.14)	65	-0.73 (0.17)	-1.31 (0.15)	99	-1.16 (0.17)	-2.27 (0.20)
32	-0.31 (0.28)	-1.01 (0.14)	66	-1.23 (0.15)	-2.25 (0.19)	100	-0.86 (0.18)	-1.13 (0.15)
33	1.22 (0.38)	0.32 (0.18)	67	-2.42 (0.37)	-1.36 (0.16)	101	-0.05 (0.16)	-0.81 (0.21)
34	-0.33 (0.32)	-0.33 (0.16)	68	-1.97 (0.47)	0.12 (0.18)			
KK ortalama SH: 0.28 BOBUT ortalama SH: 0.18								

Tablo 35 incelendiğinde, sayısal alt test kağıt-kalem uygulamasından kestirilen yetenek parametrelerinin -3.07 ile 1.47 arasında standart hata değerlerinin ise 0.00 ile 0.65 arasında değişkenlik gösterdiği görülmektedir. Sayısal alt test canlı BOBUT uygulamasından elde edilen yetenek parametreleri ise -2.92 ile 1.27 arasında standart hata değerlerinin ise 0.13 ile 0.43 arasında değişkenlik gösterdiği görülmektedir. Sözel alt teste ilişkin sonuçlar ise Tablo 36'da yer almaktadır.

Tablo 38

Katılımcıların ALES Sözel Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri θ Düzeyleri ve Standart Hataları (SH)

Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)
1	-1.63 (0.39)	-2.35 (0.31)	35	0.36 (0.09)	-1.51 (0.26)	69	-1.34 (0.45)	-1.56 (0.23)
2	1.80 (0.41)	-2.25 (0.27)	36	-0.15 (0.57)	-1.59 (0.24)	70	-0.36 (0.12)	-0.72 (0.25)
3	-0.01 (0.52)	-1.45 (0.25)	37	-0.36 (0.34)	-1.09 (0.26)	71	-1.77 (0.50)	-1.22 (0.23)
4	0.24 (0.11)	-0.50 (0.35)	38	-0.97 (0.59)	-0.91 (0.24)	72	-1.42 (0.34)	-2.52 (0.29)
5	0.09 (0.25)	-1.31 (0.24)	39	-1.10 (0.69)	-1.60 (0.25)	73	0.17 (0.08)	0.34 (0.33)
6	-0.51 (0.32)	-0.83 (0.23)	40	-0.96 (0.69)	-1.98 (0.25)	74	0.04 (0.30)	-0.99 (0.24)
7	0.59 (0.04)	-0.44 (0.24)	41	-0.53 (0.55)	-2.73 (0.34)	75	0.61 (0.19)	-0.88 (0.24)
8	-0.23 (0.17)	-1.86 (0.36)	42	-0.87 (0.50)	-2.98 (0.32)	76	0.49 (0.31)	-0.43 (0.31)
9	-0.32 (0.18)	-0.43 (0.24)	43	0.25 (0.48)	-2.52 (0.30)	77	1.91 (0.43)	0.73 (0.27)
10	-0.07 (0.15)	-0.45 (0.26)	44	-0.79 (0.29)	-1.84 (0.27)	78	0.57 (0.26)	-0.50 (0.26)
11	-0.57 (0.53)	-1.58 (0.25)	45	-2.62 (0.51)	-1.72 (0.29)	79	0.09 (0.13)	-1.24 (0.24)
12	-0.30 (0.20)	-0.94 (0.23)	46	-0.43 (0.16)	-2.27 (0.30)	80	0.15 (0.10)	0.01 (0.31)
13	1.87 (0.46)	0.34 (0.32)	47	-0.17 (0.18)	-0.99 (0.32)	81	0.36 (0.09)	0.72 (0.28)
14	-0.44 (0.21)	-0.98 (0.22)	48	-0.38 (0.17)	-1.08 (0.25)	82	0.09 (0.17)	-0.49 (0.30)
15	0.91 (0.10)	-0.43 (0.25)	49	0.43 (0.15)	-1.71 (0.26)	83	0.85 (0.20)	0.63 (0.21)
16	-0.09 (0.29)	-1.27 (0.26)	50	-0.90 (0.40)	-2.69 (0.32)	84	0.63 (0.13)	-0.01 (0.24)
17	-0.04 (0.09)	-1.23 (0.23)	51	-0.03 (0.10)	-1.69 (0.24)	85	1.66 (0.31)	0.62 (0.23)
18	0.33 (0.14)	-1.28 (0.27)	52	-0.36 (0.14)	-1.29 (0.24)	86	0.13 (0.11)	0.15 (0.28)
19	0.20 (0.03)	-0.42 (0.26)	53	-0.66 (0.32)	-1.65 (0.46)	87	0.65 (0.16)	-0.41 (0.24)
20	0.83 (0.14)	-0.41 (0.27)	54	-0.28 (0.12)	-1.50 (0.28)	88	0.56 (0.14)	0.95 (0.27)
21	-0.18 (0.14)	-1.51 (0.24)	55	0.40 (0.16)	-2.63 (0.31)	89	0.94 (0.20)	-0.23 (0.26)
22	-0.32 (0.29)	-1.47 (0.27)	56	-0.80 (0.00)	-1.12 (0.29)	90	-0.18 (0.15)	-0.37 (0.26)

(Devam ediyor)

Tablo 36 (Devam)

Katılımcıların ALES Sözel Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri θ Düzeyleri ve Standart Hataları (SH)

Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)	Birey no	KK θ (SH)	BOBUT θ (SH)
23	-2.16 (0.54)	-2.33 (0.27)	57	0.18 (0.06)	-0.91 (0.23)	91	1.03 (0.19)	0.32 (0.32)
24	-0.96 (0.69)	-1.73 (0.27)	58	-0.66 (0.34)	-1.73 (0.24)	92	0.38 (0.08)	-0.50 (0.46)
25	-0.82 (0.25)	-1.54 (0.26)	59	-0.65 (0.40)	-0.53 (0.53)	93	0.52 (0.24)	-0.53 (0.24)
26	-0.58 (0.47)	-1.36 (0.22)	60	-1.46 (0.25)	-1.98 (0.26)	94	0.22 (0.15)	-1.22 (0.26)
27	-0.21 (0.37)	-0.88 (0.23)	61	-1.51 (0.53)	-2.56 (0.31)	95	-0.03 (0.07)	-1.24 (0.24)
28	-0.33 (0.39)	0.23 (0.27)	62	-0.71 (0.14)	-1.20 (0.23)	96	-0.21 (0.05)	-1.25 (0.23)
29	-1.94 (0.57)	-0.98 (0.27)	63	0.74 (0.45)	-0.80 (0.23)	97	-0.92 (0.39)	-2.21 (0.38)
30	-1.60 (0.40)	-1.56 (0.24)	64	-2.05 (0.53)	-1.89 (0.27)	98	-2.09 (0.53)	-1.90 (0.26)
31	-1.76 (0.47)	-1.93 (0.26)	65	-0.74 (0.16)	-1.41 (0.22)	99	-0.06 (0.11)	-1.38 (0.26)
32	-0.79 (0.74)	-0.42 (0.25)	66	-2.19 (0.56)	-2.09 (0.26)	100	-0.30 (0.13)	-1.68 (0.27)
33	-1.62 (0.46)	-2.19 (0.29)	67	-1.16 (0.28)	-3.16 (0.31)	101	-0.75 (0.18)	-2.98 (0.33)
34	-0.99 (0.46)	-0.85 (0.22)	68	-0.79 (0.40)	-0.76 (0.29)			
KK ortalama SH: 0.29 BOBUT ortalama SH: 0.27								

Tablo 36 incelendiğinde, sözel alt test kağıt-kalem uygulamasından kestirilen yetenek parametrelerinin -2.62 ile 1.91 arasında standart hata değerlerinin ise 0.00 ile 0.74 arasında değişkenlik gösterdiği görülmektedir. Sözel alt test canlı BOBUT uygulamasından elde edilen yetenek parametreleri ise -3.16 ile 0.95 arasında standart hata değerlerinin ise 0.21 ile 0.53 arasında değişkenlik gösterdiği görülmektedir.

Sayısal alt teste ilişkin ortalama standart hata değerleri incelendiğinde, kağıt-kalem uygulamasında ortalama SH = 0.28, BOBUT uygulamasında ise ortalama SH = 0.18 olarak bulunmuştur. Bu fark, BOBUT uygulamasında yetenek kestirimlerinin daha düşük hata ile gerçekleştirildiğini ve dolayısıyla yetenek kestiriminin daha hassas olduğunu göstermektedir. Benzer biçimde, sözel alt testte kağıt-kalem uygulamasında ortalama SH = 0.29, BOBUT uygulamasında ise ortalama SH = 0.27 olarak elde edilmiştir. Bu fark görece küçük olmakla birlikte, BOBUT uygulamasının sözel alt testte

de biraz daha düşük hata ile kestirim yaptığını söylenebilir. Genel olarak, her iki alt testte de BOBUT kağıt-kalem uygulamasına kıyasla daha hassas yetenek kestirim yapmaktadır. Nitekim literatürde de BOBUT uygulamalarının bireylerin yetenek düzeylerine uygun madde seçimi yoluyla ölçme hassasiyetini artırdığı belirtilmektedir (Hogan, 2014; Žitný, 2012).

ALES sayısal alt test canlı BOBUT ve KK testi sonuçlarından elde edilen yetenek parametrelerinin ortalamaları arasındaki farkın manidarlığının incelendiği t-testi sonuçları Tablo 37'de verilmiştir.

Tablo 39

Katılımcıların ALES Sayısal Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri Yetenek Parametrelerine İlişkin t Testi Sonuçları

Uygulama Türü	N	Ortalama	Standart Sapma (SS)	sd	T	p
Canlı BOBUT	101	-0.516	0.841	100	1.130	0.261
KK Testi	101	-0.443	0.954	100		

Tablo 37 incelendiğinde, katılımcıların ALES sayısal alt test canlı BOBUT uygulamasından elde ettikleri yetenek puanları ile ALES kağıt-kalem testinden kestirilen yetenek düzeyleri arasında istatistiksel olarak anlamlı bir fark bulunmamıştır ($t(100)=1.130$; $p>0.05$). Bu sonuç, her iki uygulamanın da bireylerin yetenek düzeylerini benzer biçimde ölçtüğünü göstermektedir. Canlı BOBUT uygulamasından elde edilen yetenek kestirimlerinin ortalaması -0.516 (SS=0.841), KK testinden elde edilen kestirimlerin ortalaması ise -0.443 (SS=0.954) olarak bulunmuştur. Bu ortalamalar birbirine oldukça yakındır; dolayısıyla BOBUT uygulaması, katılımcıların yetenek düzeylerini, yaklaşık %60 daha az madde kullanarak ve daha düşük standart hatalarla, kağıt-kalem testinden elde edilen yetenek düzeyleriyle benzer biçimde kestirmektedir. Taylor vd (1998) tarafından yapılan çalışmada, 1996 TOEFL uygulamasında kağıt-kalem ve BOBUT uygulamalarının karşılaştırılabilirliğini incelemiş ve farklı uygulamaları alan katılımcılar arasında anlamlı bir başarı farkı bulunmamıştır. Benzer şekilde, Puhan vd (2005) tarafından yapılan bir çalışmada, öğretmenlik yeterlik testinin kağıt-kalem ve BOBUT formatlarında uygulanabilirliği incelenmiş; okuma, yazma ve matematik testlerinden elde edilen puanlar arasında anlamlı bir fark olmadığı tespit edilmiştir. Van

der Linden ve Glas (2010) tarafından yürütülen bir araştırmada ise BOBUT uygulamalarının bireylerin yetenek kestirimlerini daha az maddeyle güvenilir bir şekilde gerçekleştirebildiğini göstermiştir. Çalışmada BOBUT uygulamalarının kağıt-kalem testleriyle benzer doğruluk seviyelerine ulaşırken, test yükünü önemli ölçüde azalttığını vurgulamaktadır. Žitný vd (2012) tarafından yürütülen çalışmada, iki bilişsel yetenek testi olan Test of Intellect Potential (TIP) ve Vienna Matrices Test (VMT), BOBUT ve kağıt-kalem formatlarında uygulanabilirlik açısından karşılaştırılmıştır. Elde edilen bulgular, BOBUT ve kağıt-kalem uygulamaları arasında performans farklarının istatistiksel olarak anlamlı olmadığını göstermiştir. Ayrıca, çalışmada BOBUT uygulamalarının, kağıt-kalem testi ile benzer yetenek kestirimlerini daha az madde kullanarak (yaklaşık %55 madde tasarrufu ile) gerçekleştirdiği rapor edilmiştir. Bu sonuçlar, mevcut çalışmanın bulgularıyla tutarlılık göstermektedir.

ALES sözel alt test canlı BOBUT ve KK testi sonuçlarından elde edilen yetenek parametrelerinin ortalamaları arasındaki farkın manidarlığının incelendiği t-testi sonuçları Tablo 38'de verilmiştir.

Tablo 40

Katılımcıların ALES Sözel Alt Test Kağıt-Kalem (KK) Formundan ve Canlı BOBUT Uygulamasından Elde Ettikleri Yetenek Parametrelerine İlişkin t Testi Sonuçları

Uygulama Türü	N	Ortalama	Standart Sapma (SS)	sd	t	p
Canlı BOBUT	101	-1.175	0.896	100	10.720	0.000
KK Testi	101	-0.306	0.900	100		

Tablo 38 incelendiğinde, katılımcıların ALES sözel alt test canlı BOBUT uygulamasından elde ettikleri yetenek puanları, ALES kağıt-kalem testinden kestirilen yetenek puanlarından anlamlı bir şekilde düşük bulunmuştur ($t(100)=10.720$; $p<0.05$). Canlı BOBUT uygulamasından elde edilen yetenek kestirimlerinin ortalaması -1.175 (SS=0.896), KK testinden elde edilen kestirimlerin ortalaması ise -0.306 (SS=0.900) olarak bulunmuştur. Her iki uygulamada da standart sapmalar birbirine yakın bu da puan dağılımlarının her iki uygulamada da benzer dağılımda olduğunu göstermektedir. ALES sözel alt testinde istatistiksel olarak anlamlı bir farkın bulunması, sözel alt testinde yer

alan maddelerin yapısı, uzun paragrafların bilgisayar ekranında okunmasındaki zorluk, ekranda harcanan sürenin fazlalığı, sayısal alt testten sonra sözel alt test çözümü nedeniyle oluşan yorgunluk, dikkat dağınıklığı, konsantrasyon kaybı veya artan bilişsel yük gibi faktörlerden kaynaklanabileceğini düşündürmektedir. Bireylerin farklı ortamlarda verdikleri tepkilerdeki farklılıklar, kağıt-kalem testlerinin daha tanıdık olmasıyla da açıklanabilir (Threlfall vd., 2007). Bilgisayar tabanlı testlerde uzun okuma metinlerinin birden fazla ekranda sunulması, aynı metinlerin tek sayfada yer aldığı kağıt-kalem testlerine kıyasla performans farklılıklarına neden olabilir. Ekranda uzun metinleri okumakta zorlanan öğrenciler, testleri kağıt-kalem ortamına göre daha zor algılayabilir veya daha yavaş yanıtlayabilirler. Bu durum, farklı uygulama ortamlarının katılımcı performansını etkileyebilmesi nedeniyle bilgisayar ve kağıt-kalem ortamında alınan puanların eşdeğer olmama olasılığını ortaya koymaktadır (Thompson vd., 2003). de Beer ve Visser (1998) tarafından yürütülen çalışmada lise son sınıf öğrencilerinin üniversiteye girişte akademik yeterliklerini değerlendirmek için kullanılan GSAT (General Scholastic Aptitude Test) testinin BOBUT ve kağıt-kalem sonuçları karşılaştırılmıştır. Araştırma bulguları, katılımcıların kağıt-kalem testinde bilgisayar tabanlı teste kıyasla daha yüksek puanlar aldığını göstermiştir. Bu fark hem sözel hem de sözel olmayan alt testlerde benzer biçimde gözlenmiştir. Sonuçlar, testin uygulandığı ortamın yani bilgisayarın, katılımcıların performansı üzerinde etkili olabileceğini ve dolayısıyla uygulama ortamının test sonuçlarını değiştirebileceğini ortaya koymuştur. Wagner vd (2022) tarafından yapılan çalışmada, 3. sınıfta Matematik ve Almanca, 8. sınıfta ise Matematik, Almanca, İngilizce ve Fransızca derslerinde uygulanan VERA başarı testlerinin kağıt-kalem ve BOBUT uygulamalarında öğrencilerin yetenek düzeyleri bakımından farklılık gösterip göstermediği incelenmiştir. Çalışma sonuçları, öğrencilerin BOBUT uygulamalarında, kağıt-kalem testine kıyasla anlamlı derecede daha düşük performans sergilediklerini ortaya koymuştur. Öğrencilerin klavye kullanımı ve ekran kaydırma gibi ek zorluklarla, öğretmen ve öğrencilerin bilgisayar ortamında karşılaşabileceği teknik sorunlardan kaynaklanan endişeler test atmosferini olumsuz etkileyebileceği bu nedenle, öğrenciler bu tür taleplere alışmadan maddeleri anlamaya ve çözmeye tam olarak odaklanamayabileceği belirtilmiştir. Bu çalışmada sözel alt test canlı BOBUT uygulamasından elde edilen bulgulara benzer şekilde, Arslan vd. (2021), Dahan-Golan vd (2018), Özbaşı (2014), Şenel (2017) ve Støle vd (2020) tarafından yürütülen araştırmalarda da kağıt-kalem testiyle canlı BOBUT uygulaması arasında manidar bir

fark bulunduğu, kağıt-kalem testlerinden elde edilen puanların, BOBUT uygulamalarından elde edilen puanlardan daha yüksek olduğu sonucuna ulaşılmıştır.

Katılımcılara ALES Canlı BOBUT Uygulamasında Uygulanan Maddelerin İçerik Dağılımlarının İncelenmesi

Araştırmanın dördüncü alt amacının (c) maddesi kapsamında, gruptaki tüm katılımcılara ALES sayısal ve sözel alt testlerde uygulanan maddelerin birey bazındaki içerik dağılımlarının yüzde ve frekansları sayısal alt test için Tablo 39'da, sözel alt test için Tablo 40'ta, tüm grubun ortalama içerik dağılımları ise Tablo 41'de gösterilmiştir.

Tablo 41

ALES Sayısal Alt Test Canlı BOBUT Uygulamasında Uygulanan Maddelerin Yüzde ve Frekansları

Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)	Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)	Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)
1	30.0 (6)	55.0 (11)	15.0 (3)	35	31.5 (6)	52.6 (10)	15.7 (3)	69	30.0 (6)	55.0 (11)	15.0 (3)
2	30.0 (6)	55.0 (11)	15.0 (3)	36	30.0 (6)	55.0 (11)	15.0 (3)	70	30.0 (6)	55.0 (11)	15.0 (3)
3	30.0 (6)	55.0 (11)	15.0 (3)	37	30.0 (6)	55.0 (11)	15.0 (3)	71	30.0 (6)	55.0 (11)	15.0 (3)
4	30.0 (6)	55.0 (11)	15.0 (3)	38	30.0 (6)	55.0 (11)	15.0 (3)	72	30.0 (6)	55.0 (11)	15.0 (3)
5	30.0 (6)	55.0 (11)	15.0 (3)	39	30.0 (6)	55.0 (11)	15.0 (3)	73	30.0 (6)	55.0 (11)	15.0 (3)
6	31.5 (6)	52.6 (10)	15.7 (3)	40	30.0 (6)	55.0 (11)	15.0 (3)	74	27.7 (5)	55.5 (10)	16.6 (3)
7	30.0 (6)	55.0 (11)	15.0 (3)	41	30.0 (6)	55.0 (11)	15.0 (3)	75	30.0 (6)	55.0 (11)	15.0 (3)
8	31.5 (6)	52.6 (10)	15.7 (3)	42	30.0 (6)	55.0 (11)	15.0 (3)	76	30.0 (6)	55.0 (11)	15.0 (3)
9	30.0 (6)	55.0 (11)	15.0 (3)	43	30.0 (6)	55.0 (11)	15.0 (3)	77	30.0 (6)	55.0 (11)	15.0 (3)
10	30.0 (6)	55.0 (11)	15.0 (3)	44	28.5 (4)	57.1 (8)	14.3 (2)	78	30.0 (6)	55.0 (11)	15.0 (3)
11	27.2 (3)	54.5 (6)	18.2 (2)	45	30.0 (6)	55.0 (11)	15.0 (3)	79	30.0 (6)	55.0 (11)	15.0 (3)
12	30.0 (6)	55.0 (11)	15.0 (3)	46	30.0 (6)	55.0 (11)	15.0 (3)	80	27.7 (5)	55.5 (10)	16.6 (3)
13	31.5 (6)	52.6 (10)	15.7 (3)	47	30.0 (6)	55.0 (11)	15.0 (3)	81	30.0 (6)	55.0 (11)	15.0 (3)

(Devam ediyor)

Tablo 39 (Devam)

ALES Sayısal Alt Test Canlı BOBUT Uygulamasında Uygulanan Maddelerin Yüzde ve Frekansları

Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)	Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)	Birey No	Sİ %(f)	SMPÇ %(f)	G %(f)
14	30.0 (6)	55.0 (11)	15.0 (3)	48	30.0 (6)	55.0 (11)	15.0 (3)	82	30.0 (6)	55.0 (11)	15.0 (3)
15	30.0 (6)	55.0 (11)	15.0 (3)	49	30.0 (6)	55.0 (11)	15.0 (3)	83	30.0 (6)	55.0 (11)	15.0 (3)
16	30.0 (6)	55.0 (11)	15.0 (3)	50	30.7 (4)	53.8 (7)	15.3 (2)	84	30.0 (6)	55.0 (11)	15.0 (3)
17	30.0 (6)	55.0 (11)	15.0 (3)	51	30.0 (6)	55.0 (11)	15.0 (3)	85	30.0 (6)	55.0 (11)	15.0 (3)
18	30.0 (6)	55.0 (11)	15.0 (3)	52	30.0 (6)	55.0 (11)	15.0 (3)	86	30.0 (6)	55.0 (11)	15.0 (3)
19	30.0 (6)	55.0 (11)	15.0 (3)	53	31.5 (6)	52.6 (10)	17.7 (3)	87	33.3 (4)	50.0 (6)	16.6 (2)
20	30.0 (6)	55.0 (11)	15.0 (3)	54	30.0 (6)	55.0 (11)	15.0 (3)	88	30.0 (6)	55.0 (11)	15.0 (3)
21	30.0 (6)	55.0 (11)	15.0 (3)	55	30.0 (6)	55.0 (11)	15.0 (3)	89	30.0 (6)	55.0 (11)	15.0 (3)
22	30.0 (6)	55.0 (11)	15.0 (3)	56	30.0 (6)	55.0 (11)	15.0 (3)	90	30.0 (6)	55.0 (11)	15.0 (3)
23	30.0 (6)	50.0 (10)	20.0 (4)	57	30.0 (6)	55.0 (11)	15.0 (3)	91	30.0 (6)	55.0 (11)	15.0 (3)
24	30.0 (6)	55.0 (11)	15.0 (3)	58	30.0 (6)	55.0 (11)	15.0 (3)	92	30.0 (6)	55.0 (11)	15.0 (3)
25	30.0 (6)	55.0 (11)	15.0 (3)	59	30.0 (6)	55.0 (11)	15.0 (3)	93	27.7 (5)	55.5 (10)	16.6 (3)
26	30.0 (6)	55.0 (11)	15.0 (3)	60	27.7 (5)	55.5 (10)	16.6 (3)	94	30.0 (6)	55.0 (11)	15.0 (3)
27	29.5 (5)	52.9 (9)	17.6 (3)	61	30.0 (6)	55.0 (11)	15.0 (3)	95	30.0 (6)	55.0 (11)	15.0 (3)
28	30.0 (6)	55.0 (11)	15.0 (3)	62	30.0 (6)	55.0 (11)	15.0 (3)	96	30.0 (6)	55.0 (11)	15.0 (3)
29	30.0 (6)	55.0 (11)	15.0 (3)	63	30.0 (6)	55.0 (11)	15.0 (3)	97	30.0 (6)	55.0 (11)	15.0 (3)
30	30.0 (6)	55.0 (11)	15.0 (3)	64	30.0 (6)	55.0 (11)	15.0 (3)	98	30.0 (6)	55.0 (11)	15.0 (3)
31	30.0 (6)	55.0 (11)	15.0 (3)	65	30.0 (6)	50.0 (10)	20.0 (4)	99	33.3 (4)	50.0 (6)	16.6 (2)
32	30.0 (6)	55.0 (11)	15.0 (3)	66	30.0 (6)	55.0 (11)	15.0 (3)	100	25.0 (5)	60.0 (12)	15.0 (3)
33	30.0 (6)	55.0 (11)	15.0 (3)	67	30.0 (6)	55.0 (11)	15.0 (3)	101	30.0 (6)	55.0 (11)	15.0 (3)
34	30.0 (6)	55.0 (11)	15.0 (3)	68	30.0 (6)	55.0 (11)	15.0 (3)				

Sİ: Sayılarla İşlem, SMPÇ: Sayısal mantık ve problem çözme, G: Geometri, N: Madde Sayısı

Tablo 39 incelendiğinde, ALES sayısal alt testine ait canlı BOBUT uygulamasında içerik dengelemesinin genel olarak başarılı bir şekilde sağlandığı görülmektedir. Katılımcıların büyük çoğunluğunun (%86) 20 madde cevapladığı dikkate alındığında, bu bireylerde uygulanan madde yüzdelerinin canlı uygulama öncesinde belirlenen içerik dağılımı olan, sayılarla işlem (Sİ) %30, sayısal mantık ve problem çözme (SMPÇ) %56 ve geometri (G) %14 değerleriyle neredeyse tamamen örtüştüğü söylenebilir. Nitekim 20 madde cevaplayan bireylerde Sİ yaklaşık 6 madde, SMPÇ 10–11 madde ve G 3 madde olarak dağılım göstermiştir. Bu durum, BOBUT uygulamasının madde seçimi sırasında içerik oranlarını koruyabildiği ve yalnızca toplam madde sayısının düştüğü durumlarda yüzdeliklerde matematiksel sapmaların ortaya çıktığı belirlenmiştir. Bazı bireyler alt test için verilen süre içerisinde 20 maddeyi cevaplayamadıkları için toplam madde sayısı 20'nin altında kalmıştır. Örneğin 11 madde cevaplayan Birey 11'de Sİ %27.2 (3 madde), SMPÇ %54.5 (6 madde) ve G %18.2 (2 madde) oranları elde edilmiştir. Bu bireyin yalnızca 2 geometri maddesi alması, toplam madde sayısının düşük olması nedeniyle G yüzdesinin hedef oran olan %14'e kıyasla daha yüksek görünmesine neden olmuştur. Benzer şekilde 14 madde cevaplayan Birey 44'te Sİ %28.5 (4), SMPÇ %57.1 (8) ve G %14.3 (2) değerleri gözlenmiş; burada da geometri maddesi sayısının 3 yerine 2 olması küçük bir yüzdelik kaymaya yol açmıştır. 13 madde yanıtlayan Birey 50'de ise Sİ %30.7 (4), SMPÇ %53.8 (7) ve G %15.3 (2) oranları ile içerik dengesinin büyük ölçüde korunduğu görülmektedir. Son olarak 12 madde yanıtlayan Birey 99'da Sİ %33.3 (4), SMPÇ %50.0 (6) ve G %16.6 (2) değerleri elde edilmiştir. Tüm katılımcılar birlikte değerlendirildiğinde, Sİ maddelerinin çoğunlukla 6 madde olarak uygulandığı, erken bitenlerde ise bunun doğal olarak 4–5 maddeye düştüğü görülmüştür. SMPÇ maddeleri hedef oran olan %56'ya uygun şekilde büyük ölçüde 10–11 madde olarak uygulanmış, erken bitenlerde ise bu sayı 6–8 aralığına gerilemiştir. Geometri maddeleri ise standart uygulamada 3 madde olarak görünmekte, 20 maddeden az uygulanan testlerde ise bu sayı 2 maddeye düşmekte ve bu durum beklenen bir örüntü sergilemektedir. Bu bulgular, erken sonlanan testlerde içerik yüzdelerindeki sapmaların gerçek içerik değişiminden değil, düşük madde sayısının oluşturduğu oransal etkiden kaynaklandığını göstermektedir.

Tablo 42

ALES Sözel Alt Test Canlı BOBUT Uygulamasında Uygulanan Maddelerin Yüzde ve Frekansları

Birey No	DBAB %(f)	P %(f)	SM %(f)	Birey No	DBAB %(f)	P %(f)	SM %(f)	Birey No	DBAB %(f)	P %(f)	SM %(f)
1	35.0 (7)	45.0 (9)	20.0 (4)	35	35.0 (7)	45.0 (9)	20.0 (4)	69	35.0 (7)	45.0 (9)	20.0 (4)
2	35.0 (7)	45.0 (9)	20.0 (4)	36	35.7 (5)	42.8 (6)	21.4 (3)	70	35.0 (7)	45.0 (9)	20.0 (4)
3	35.0 (7)	45.0 (9)	20.0 (4)	37	35.0 (7)	45.0 (9)	20.0 (4)	71	35.0 (7)	45.0 (9)	20.0 (4)
4	35.0 (7)	45.0 (9)	20.0 (4)	38	35.0 (7)	45.0 (9)	20.0 (4)	72	35.0 (7)	45.0 (9)	20.0 (4)
5	35.0 (7)	45.0 (9)	20.0 (4)	39	35.0 (7)	45.0 (9)	20.0 (4)	73	35.0 (7)	45.0 (9)	20.0 (4)
6	35.0 (7)	45.0 (9)	20.0 (4)	40	35.0 (7)	45.0 (9)	20.0 (4)	74	35.0 (7)	45.0 (9)	20.0 (4)
7	35.0 (7)	45.0 (9)	20.0 (4)	41	35.0 (7)	45.0 (9)	20.0 (4)	75	35.0 (7)	45.0 (9)	20.0 (4)
8	35.0 (7)	45.0 (9)	20.0 (4)	42	35.0 (7)	45.0 (9)	20.0 (4)	76	35.3 (6)	47.0 (8)	17.6 (3)
9	35.0 (7)	45.0 (9)	20.0 (4)	43	35.0 (7)	45.0 (9)	20.0 (4)	77	35.0 (7)	45.0 (9)	20.0 (4)
10	35.0 (7)	45.0 (9)	20.0 (4)	44	35.0 (7)	45.0 (9)	20.0 (4)	78	35.0 (7)	45.0 (9)	20.0 (4)
11	35.0 (7)	45.0 (9)	20.0 (4)	45	35.0 (7)	45.0 (9)	20.0 (4)	79	35.0 (7)	45.0 (9)	20.0 (4)
12	35.0 (7)	45.0 (9)	20.0 (4)	46	35.0 (7)	45.0 (9)	20.0 (4)	80	35.0 (7)	45.0 (9)	20.0 (4)
13	35.0 (7)	45.0 (9)	20.0 (4)	47	35.0 (7)	45.0 (9)	20.0 (4)	81	35.0 (7)	45.0 (9)	20.0 (4)
14	35.0 (7)	45.0 (9)	20.0 (4)	48	35.0 (7)	45.0 (9)	20.0 (4)	82	35.0 (7)	45.0 (9)	20.0 (4)
15	35.0 (7)	45.0 (9)	20.0 (4)	49	35.0 (7)	45.0 (9)	20.0 (4)	83	35.0 (7)	45.0 (9)	20.0 (4)
16	35.0 (7)	45.0 (9)	20.0 (4)	50	35.0 (7)	45.0 (9)	20.0 (4)	84	35.0 (7)	45.0 (9)	20.0 (4)
17	35.0 (7)	45.0 (9)	20.0 (4)	51	35.0 (7)	45.0 (9)	20.0 (4)	85	35.0 (7)	45.0 (9)	20.0 (4)
18	35.0 (7)	45.0 (9)	20.0 (4)	52	35.0 (7)	45.0 (9)	20.0 (4)	86	35.0 (7)	45.0 (9)	20.0 (4)
19	35.0 (7)	45.0 (9)	20.0 (4)	53	35.0 (7)	45.0 (9)	20.0 (4)	87	35.0 (7)	45.0 (9)	20.0 (4)
20	35.0 (7)	45.0 (9)	20.0 (4)	54	35.0 (7)	45.0 (9)	20.0 (4)	88	35.0 (7)	45.0 (9)	20.0 (4)
21	35.0 (7)	45.0 (9)	20.0 (4)	55	35.0 (7)	45.0 (9)	20.0 (4)	89	35.0 (7)	45.0 (9)	20.0 (4)
22	35.0 (7)	45.0 (9)	20.0 (4)	56	35.0 (7)	45.0 (9)	20.0 (4)	90	35.0 (7)	45.0 (9)	20.0 (4)

(Devam ediyor)

Tablo 40 (Devam)

ALES Sözel Alt Test Canlı BOBUT Uygulamasında Uygulanan Maddelerin Yüzde ve Frekansları

Birey No	DBAB %(f)	P %(f)	SM %(f)	Birey No	DBAB %(f)	P %(f)	SM %(f)	Birey No	DBAB %(f)	P %(f)	SM %(f)
23	35.0 (7)	45.0 (9)	20.0 (4)	57	35.0 (7)	45.0 (9)	20.0 (4)	91	35.0 (7)	45.0 (9)	20.0 (4)
24	35.0 (7)	45.0 (9)	20.0 (4)	58	35.0 (7)	45.0 (9)	20.0 (4)	92	35.0 (7)	45.0 (9)	20.0 (4)
25	35.0 (7)	45.0 (9)	20.0 (4)	59	35.0 (7)	45.0 (9)	20.0 (4)	93	35.0 (7)	45.0 (9)	20.0 (4)
26	35.0 (7)	45.0 (9)	20.0 (4)	60	35.0 (7)	45.0 (9)	20.0 (4)	94	35.0 (7)	45.0 (9)	20.0 (4)
27	35.0 (7)	45.0 (9)	20.0 (4)	61	35.0 (7)	45.0 (9)	20.0 (4)	95	35.0 (7)	45.0 (9)	20.0 (4)
28	40.0 (8)	40.0 (8)	20.0 (4)	62	35.0 (7)	45.0 (9)	20.0 (4)	96	35.0 (7)	45.0 (9)	20.0 (4)
29	40.0 (8)	40.0 (8)	20.0 (4)	63	30.0 (6)	50.0 (10)	20.0 (4)	97	35.0 (7)	45.0 (9)	20.0 (4)
30	40.0 (8)	40.0 (8)	20.0 (4)	64	35.0 (7)	45.0 (9)	20.0 (4)	98	35.0 (7)	45.0 (9)	20.0 (4)
31	35.0 (7)	45.0 (9)	20.0 (4)	65	35.0 (7)	45.0 (9)	20.0 (4)	99	35.0 (7)	45.0 (9)	20.0 (4)
32	35.0 (7)	45.0 (9)	20.0 (4)	66	35.0 (7)	45.0 (9)	20.0 (4)	100	35.0 (7)	45.0 (9)	20.0 (4)
33	35.0 (7)	45.0 (9)	20.0 (4)	67	35.0 (7)	45.0 (9)	20.0 (4)	101	35.0 (7)	45.0 (9)	20.0 (4)
34	35.0 (7)	45.0 (9)	20.0 (4)	68	35.0 (7)	45.0 (9)	20.0 (4)				

DBAB: Dil Bilgisi ve Anlam Bütünlüğü, P: Paragraf, SM: Sözel Mantık, N: Madde Sayısı

Tablo 40 incelendiğinde, ALES sözel alt testine ait canlı BOBUT uygulamasında da içerik dengelemesinin tutarlı bir biçimde sağlandığı görülmektedir. Katılımcıların büyük çoğunluğunun (%99) 20 madde uygulaması aldığı göz önünde bulundurulduğunda, bu bireylerde uygulanan madde yüzdelerinin canlı uygulama öncesinde belirlenen içerik dağılımı olan Dil Bilgisi ve Anlam Bütünlüğü (DBAB) %36, Paragraf (P) %47 ve Sözel Mantık (SM) %17 hedef oranlarıyla çok yakın olduğu söylenebilir. 20 madde yanıtlayan katılımcıların büyük kısmında içerik dağılımı DBAB 7 madde, Paragraf 9 madde ve Sözel Mantık 4 madde olacak şekilde standart bir örüntü sergilemektedir. Bununla birlikte bazı katılımcılar testte yer alan 20 maddeyi cevaplamasına rağmen içerik oranlarında küçük matematiksel sapmalar gözlenmiştir. Örneğin, birey 28, birey 29 ve birey 30 gibi bazı katılımcılarda DBAB ve Paragraf maddeleri 8'er madde (%40–%40) olacak şekilde uygulanmış, ancak Sözel Mantık 4 madde ile aynı kalmıştır. Bu tür küçük değişimlerin, bireyin geçici yetenek düzeyine uygun olacak şekilde madde seçim algoritmasının, bilgi

fonksiyonuna dayanarak yaptığı seçimlerden kaynaklandığı ve içerik dengesinin genel işleyişini bozmadığı görülmektedir. Benzer şekilde Birey 63'te DBAB maddeleri 6 madde (%30) ve Paragraf maddeleri 10 madde (%50) şeklinde gözlenmiştir. Burada da toplam madde sayısı aynı olmakla birlikte içerik dağılımı hedef oranlara oldukça yakın olup, farklılığın temel nedeni adaptif algoritmanın bireye uygun maddeleri seçmesidir. Bu değerler, belirlenen hedef oranlarla uyumludur ve BOBUT uygulamasının madde seçiminde içerik ağırlıklarını başarıyla koruduğunu göstermektedir.

Tablo 43

ALES Alt Testlerinde Canlı BOBUT Uygulamasında Tüm Grup İçin Uygulanan Maddelerin Ortalama İçerik Dağılımları

Alt Test / İçerik	Uygulanan Ort. Madde Sayısı (N)	Canlı BOBUT İçerik Yüzdesi (%)	KK İçerik Yüzdesi (%)
SAYISAL			
Sayılarla İşlem	5.8	30.0	30.0
Sayısal mantık ve problem çözme	10.7	54.7	56.0
Geometri	3.0	15.3	14.0
Sayısal Alt Test Toplam	20	100.0	100.0
SÖZEL			
Dil Bilgisi ve Anlam Bütünlüğü	7.0	35.1	36.0
Paragraf	8.9	44.9	47.0
Sözel mantık	4.0	20.0	17.0
Sözel Alt Test Toplam	20	100.0	100.0

ALES alt testlerinde Canlı BOBUT uygulaması sırasında tüm grup için uygulanan maddelerin içerik dağılımlarının yer aldığı Tablo 41 incelendiğinde, BOBUT uygulamasının KK testine oldukça yakın bir içerik dengesi sağladığı görülmektedir. Sayısal alt testte, sayılarla işlem (%30), sayısal mantık ve problem çözme (%54.7) ile geometri (%15.3) içerik oranları, KK testindeki karşılıklarıyla (%30, %56 ve %14) büyük ölçüde uyum göstermektedir. Benzer şekilde, sözel alt testte dil bilgisi ve anlam bütünlüğü (%35.1), paragraf (%44.9) ve sözel mantık (%20) içerik oranları da KK formuna (%36, %47 ve %17) oldukça yakın bulunmuştur.

Özellikle birey hakkında kritik kararların verileceği yüksek riskli sınavlarda, sınav ile içerik düzeyleri arasında yüksek düzeyde uyum sağlanması, eğitimde hesap verebilirliğin önemli bir göstergesi olarak öne çıkmaktadır. Uyarlanmış testlerde içerik uyumunun değerlendirilmesine yönelik araştırmalar sınırlı olmakla birlikte (Wise vd.,

2015), mevcut çalışmalar (Öcal ve Doğan, 2024; Sarı ve Manley, 2017; Shin vd., 2009; Yasuda ve Hull, 2021; Zheng vd., 2013) genellikle farklı örneklem büyüklükleri, içerik sayıları ve madde seçim yöntemleri, yetenek kestirim yöntemleri altında simülasyon çalışmaları biçiminde yürütülmektedir. Bu çalışmaların çoğunda, kapsam dengelemenin madde havuzunun verimli kullanımını sağlamada, madde kullanım sıklığını azaltmada ve ölçme kesinliğini korumada etkili olduğu belirlenmiştir. Gerçek BOBUT uygulamalarında ise içerik uyumunun değerlendirilmesi genellikle sınav uygulayıcıları ve alan uzmanları tarafından gerçekleştirilmektedir. Bu alandaki araştırmaların artmasıyla birlikte, içerik uyumu değerlendirme kriterlerinin de daha kapsamlı ve sistematik hâle gelmesi beklenmektedir (Wise vd., 2015).

BÖLÜM 5

SONUÇ VE ÖNERİLER

Bu bölümde araştırma bulgularına dayalı olarak ulaşılan sonuçlara ve bu sonuçlara dayalı olarak önerilere yer verilmiştir.

Sonuçlar

Bu araştırmada ALES'in BOBUT olarak uygulanabilirliği incelenmiştir. Araştırmada, ÖSYM'den temin edilen veriler kullanılarak ALES'in sayısal ve sözel alt testleri için ayrı ayrı madde ve yetenek kestirimleri yapılmış ve aynı veri seti üzerinden post hoc simülasyonlar gerçekleştirilmiştir. Araştırma kapsamında, farklı yetenek kestirim, madde seçimi ve test sonlandırma yöntemlerinin kombinasyonlarından oluşan 40 farklı koşul altında post hoc simülasyonlar yürütülmüştür. Elde edilen bulgular doğrultusunda en uygun BOBUT koşulları belirlenmiş ve bu koşullar dikkate alınarak ALES canlı BOBUT uygulaması geliştirilmiştir. Geliştirilen ALES canlı BOBUT ile ALES kâğıt-kalem testi 101 kişiden oluşan çalışma grubuna bir hafta arayla uygulanmış ve iki uygulamadan elde edilen sonuçlar karşılaştırılmıştır. Bulgular ışığında ulaşılan sonuçlar aşağıda maddeler hâlinde sunulmuştur.

1. ALES'te yer alan maddelerin ikili puanlanma özelliği dikkate alınarak, parametre kestirimlerinde 2 ve 3 parametreli lojistik modellerin model veri uyumları incelenmiş ve en uygun modelin 3PL olduğu sonucuna ulaşılmıştır.
2. ALES sayısal alt testinde yer alan 150 maddenin tamamının ayırt edicilik ve şans parametreleri bakımından kabul edilebilir aralıkta olduğu görülmüştür. Güçlük parametreleri açısından değerlendirildiğinde ise, havuzdaki maddelerin büyük çoğunluğunun orta ve kolay düzeyde olduğu, zor düzeydeki maddelerin ise sınırlı sayıda bulunduğu sonucuna ulaşılmıştır. ALES sözel alt testinde yer alan 150 maddeden 9'unun, ayırt edicilik parametresinin yetersiz bulunması ve doğrulayıcı faktör analizi sonuçlarının uygun olmaması nedeniyle madde havuzundan çıkarılmıştır. Geriye kalan 141 maddeden bazıları şans parametresi açısından kabul edilebilir sınırların üzerinde olsa da yüksek ayırt edicilik değerleri nedeniyle madde havuzunda tutulmuştur. Güçlük parametreleri bakımından

değerlendirildiğinde ise, madde havuzunun kolay, orta ve zor düzeydeki maddelerden dengeli bir şekilde oluştuğu sonucuna ulaşılmıştır.

3. Post hoc simülasyonları sonucunda hem sayısal hem de sözel alt testler için en uygun BOBUT koşulunun; "Beklenen Sonsal Dağılım" yetenek kestirim yönteminin, "sabit madde sayısı MS = 20" test sonlandırma kuralının ve "Maksimum Fisher Bilgisi" madde seçim yönteminin olduğu belirlenmiştir. Post hoc simülasyonları sonucunda alt testler 20 madde ile sonlandırıldığında test uzunluğu her iki alt test için %60 oranında azalmıştır.
4. Katılımcıların ALES sayısal alt test canlı BOBUT ve kağıt-kalem formlarından elde edilen yetenek kestirimleri arasında istatistiksel olarak manidar bir fark bulunmamıştır. Bu durum, ALES sayısal alt testinin BOBUT olarak uygulanması halinde %60 oranında daha az madde ile kağıt-kalem testine benzer şekilde yetenek kestirimi yapabileceğini göstermektedir.
5. Katılımcıların ALES sözel alt testte canlı BOBUT ve kağıt-kalem formlarından elde edilen yetenek kestirimleri arasında istatistiksel olarak manidar bir fark bulunmuştur ve bu fark kağıt-kalem testi lehinedir. Bu durum, sözel alt testte madde yapısı, içerik özellikleri ve ek bilişsel yük gibi dijital test ortamının etkileri nedeniyle BOBUT uygulamasının sayısal testte olduğu kadar tutarlı sonuç üretmediğine düşündürmektedir.
6. ALES'in kâğıt–kalem ve BOBUT uygulamalarından elde edilen yetenek kestirimleri incelendiğinde, sözel alt test için .59, sayısal alt test için ise .75 olmak üzere orta ve yüksek düzeyde korelasyon değerleri elde edilmiştir. Korelasyonun sayısal alt testte daha yüksek olması, bu alt testin iki uygulama arasında daha tutarlı ve örtüşen yetenek kestirimleri ürettiğini göstermektedir. Sözel alt testteki .59 korelasyon ise uygulamaların bireyleri kısmen benzer bir biçimde sıraladığını ortaya koymaktadır. Bununla birlikte, sözel alt testte BOBUT ve kâğıt–kalem uygulamaları arasında istatistiksel olarak manidar bir fark bulunması ve bu farkın kâğıt–kalem lehine olması, bireylerin kâğıt–kalem formunda BOBUT'a kıyasla daha yüksek yetenek düzeyleri elde ettiğini göstermektedir.
7. ALES alt testlerinde canlı BOBUT uygulaması sırasında tüm grup için seçilen maddelerin içerik dağılımları incelendiğinde, BOBUT'un kâğıt–kalem formuna oldukça yakın bir içerik dengesi sağladığı görülmüştür. Sayısal alt testte sayılarla işlem, sayısal mantık ve problem çözme ile geometrinin; sözel alt testte ise dil bilgisi ve anlam bütünlüğü, paragraf ve sözel mantık içeriklerinin BOBUT

uygulamasındaki oranlarının kâğıt–kalem formundaki dağılımlarla büyük ölçüde örtüştüğü belirlenmiştir. Bu sonuçlar, BOBUT'un hem sayısal hem de sözel alt testlerde madde seçimini içerik alanlarına göre dengeli biçimde gerçekleştirdiğini ve ALES'in içerik yapısını başarıyla koruduğunu göstermektedir.

Öneriler

Araştırma bulguları ile araştırma sürecinde karşılaşılan güçlükler birlikte değerlendirildiğinde, gelecekte yapılacak çalışmalar ve benzer konularda çalışacak araştırmacılar için aşağıdaki önerilerin sunulması önemli görülmektedir.

Araştırmacılara Yönelik Öneriler

1. Bu araştırmada canlı BOBUT uygulamasında içerik dengelemesi kullanılmış ve test sürecinin içerik açısından geçerli olması sağlanmıştır. Bununla birlikte, içerik dengelemesinin adaptif test süreçlerinde ölçme hassasiyetini nasıl etkilediği alan yazında halen tartışılan bir konudur. Bu nedenle ileride yapılacak çalışmalarda, içerik dengelemesi yapılmış ve yapılmamış canlı BOBUT uygulamalarının karşılaştırılması önem taşımaktadır. Bu doğrultuda, içerik dengelemesi yapılan ve yapılmayan canlı BOBUT uygulamalarının karşılaştırılmasına yönelik çalışmaların hem ALES gibi geniş ölçekli sınavlarda hem de düşük riskli eğitim ortamlarında ölçme hassasiyeti açısından önemli sonuçlar ortaya koyacağı düşünülmektedir.
2. Bu araştırmada sayısal alt test canlı BOBUT uygulamasında ortalama standart hatanın sözel alt test canlı BOBUT uygulamasına kıyasla daha düşük çıkması, büyük ölçüde sayısal alt test madde havuzunda ayırt ediciliği yüksek maddelerin çok daha yüksek oranda (%80) yer almasına karşın, sözel alt testte yüksek ayırt ediciliğe sahip maddelerin oranının görece düşük (%27.6) olmasından kaynaklanmaktadır. Bu nedenle, ileride yapılacak BOBUT uygulamalarında yetenek kestirimlerinin daha hassas yapılabilmesi için, sayısal alt testte olduğu gibi ayırt ediciliği yüksek maddelerin madde havuzundaki oranının artırılması önerilmektedir.
3. Bu çalışmada, ALES sözel alt testinin kâğıt-kalem uygulamasından elde edilen yetenek düzeylerinin canlı BOBUT uygulamasına göre daha yüksek olduğu görülmüştür. Bu farkın, özellikle sözel testte adayların sürekli yeni metinlerle karşılaşması nedeniyle artan ekran süresi, bilişsel yük ve bağlam değişiminin

performansı olumsuz etkilemesinden kaynaklanabileceği düşünülmektedir. ALES'in ortak köklü metin temelli yapısı, canlı BOBUT uygulamalarında adayların 8–10 farklı metin okumak zorunda kalmasına yol açarak aday deneyimini sınırlayabilmekte ve gerçek performansın tam olarak yansıtılamamasına neden olabilmektedir. Bu nedenle mevcut BOBUT yaklaşımının, modül bazlı bir yapı sunan Çok Aşamalı Bireye Uyarlanmış Test (ÇABOBUT; Multistage Adaptive Testing, MST) modeliyle karşılaştırılması önerilmektedir. Bu bağlamda, testlet (madde takımı) temelli MST modeli ile madde bazlı BOBUT yaklaşımının ölçme hassasiyeti, yetenek kestirim düzeyleri, aday deneyimi ve bilişsel yük açısından karşılaştırılması hem gelecekteki araştırmalar hem de ALES'in operasyonel süreçlerinin geliştirilmesi için önemli katkılar sağlayacaktır.

4. Bu çalışmada kâğıt-kalem ve canlı BOBUT uygulamaları arasında bazı farklılıklar gözlenmiştir. Bu durum, bilgisayar tabanlı ölçme ortamlarının aday performansını çeşitli açılardan etkileyebileceğine işaret etmektedir. Elde edilen sonuçların altında yatan nedenleri daha derinlemesine anlamak ve aday deneyimini bütüncül olarak değerlendirebilmek için gelecekte canlı BOBUT uygulamalarının nitel araştırmalarla desteklenmesi önerilmektedir.
5. Bu araştırmada canlı BOBUT uygulamasında katılımcılar sırasıyla sayısal ve ardından sözel alt testi çözmüştür. Oysa kağıt-kalem uygulamasında katılımcılar, istedikleri alt testten başlayabilme esnekliğine sahiptir. Bu farklılığın giderilmesi ve test deneyiminin iyileştirilmesi amacıyla, canlı BOBUT uygulamalarının yapılacağı çalışmalarda katılımcılara istediği alt testten başlama seçeneğinin sunulması önerilmektedir.
6. Araştırma bulguları, Ağırlıklandırılmış Olabilirlik Bilgi (AOB) madde seçim yönteminin post hoc simülasyonlarında Maksimum Fisher bilgisine benzer performans gösterdiğini ortaya koymaktadır. Bu nedenle, ileriye dönük çalışmalarda AOB madde seçim yönteminin farklı post hoc koşullarında detaylı olarak incelenmesi ve ölçme hassasiyeti açısından değerlendirilmesi önerilmektedir. Eğer sonuçlar olumlu olursa, bu yöntem canlı BOBUT uygulamalarında da kullanılabilir.
7. Farklı öğrenci grupları, üniversite sınıf düzeyleri ve eğitim geçmişine sahip katılımcılar üzerinde benzer karşılaştırmalar yapılması, BOBUT uygulamalarının genellenebilirliğini artıracağı düşünülmektedir.

Uygulamaya Yönelik Öneriler

1. Bu çalışmada katılımcıların önemli bir kısmının daha önce bilgisayar ortamında sınava girmediği ve ilk defa bilgisayar ortamında sınav deneyimi yaşadıkları uygulama sürecinde gözlemlenmiştir. Bu durumun test performansını dolaylı olarak etkileyebileceği dikkate alındığında, BOBUT uygulamalarının güvenilirliğini artırmak için adayların bilgisayar tabanlı sınavlara önceden hazırlanması, temel bilgisayar aşinalığının desteklenmesi ve test arayüzüne yönelik kısa bir tanıtım sürecinin sağlanması önerilmektedir. Böyle bir hazırlık süreci, test kaygısını azaltarak adayların gerçek performanslarının ölçülmesine katkı sağlayacaktır. Ayrıca uygulayıcıların bilgisayar tabanlı test süreçlerine yönelik teknik bilgi ve becerilerinin güçlendirilmesi, olası aksaklıkların önüne geçerek ölçme sürecinin bütünlüğünü destekleyecektir.
2. Bilgisayar tabanlı sınavların adil ve geçerli biçimde uygulanabilmesi için adayların bilgisayar aşinalığı, ekran üzerinden okuma becerileri ve bilişsel yük düzeyleri gibi kullanıcı özelliklerinin sınav performansı üzerindeki etkilerinin sistematik biçimde değerlendirilmesi gerekmektedir. Bu doğrultuda ÖSYM'nin, sınav öncesi ve sonrası süreçleri kapsayan bir pilot kullanıcı deneyimi uygulaması yürütmesi önerilmektedir. Bu pilot uygulama; adayların bilgisayar kullanma sıklığı, ekran metni okuma stratejileri, dijital ortamlarda soru çözme davranışları, uzun metinlerde yaşanan yorgunluk durumu ve bilgisayar tabanlı testlerde karşılaştıkları bilişsel zorlukları ortaya koyabilir. Elde edilecek bulgular doğrultusunda sınav arayüzü tasarımının iyileştirilmesi, font ve ekran düzenlemelerinin optimize edilmesi, uzun metinli maddeler için daha uygun sunum biçimlerinin geliştirilmesi ve adayların dijital sınav ortamına uyumunu artıracak yönlendirmelerin yapılması mümkün olacaktır.
3. Bilgisayar tabanlı sınavlarda uzun oturum süreleri, özellikle metin ağırlıklı sözel testlerde adaylarda dikkat azalmasına ve test yorgunluğuna yol açabilmektedir. Bu durum performansın düşmesine neden olduğundan, ölçmenin geçerliğini de olumsuz etkileyebilir. Bu nedenle ÖSYM'nin, bilgisayar tabanlı sınavlarda bölümlendirilmiş oturum yapıları veya adayın dikkatini toparlamasına imkân veren kısa molalar gibi uygulamaları değerlendirmesi önerilmektedir. Böyle uygulamalar, adayın bilişsel yükünü dengeleyerek sınav süresince daha istikrarlı bir performans sergilemesine katkı sağlayacaktır.

4. Çalışmada kullanılan madde havuzunda, özellikle yüksek yetenek düzeylerini temsil eden maddelerin sınırlı olduğu görülmüştür. Sayısal ve sözel alt testlerdeki güçlük parametresi ortalamaları, madde havuzlarının daha çok düşük ve orta yetenek düzeylerine uygun maddelerden oluştuğunu göstermektedir. Bu durum, canlı BOBUT uygulamasında yüksek yetenek düzeyindeki bireylere uygun güçlükte madde seçimini zorlaştırmış ve bu bireylerin yetenek kestirimlerinin daha yüksek hata ile kestirilmesine neden olmuştur. Bilgisayar tabanlı bireye uyarlanmış test uygulamalarının yaygınlaştırılması hedeflendiğinde, madde havuzunun tüm yetenek düzeylerini kapsayacak biçimde geniş bir güçlük dağılımına sahip olması büyük önem taşımaktadır.

KAYNAKLAR

Alkan, V. (2021). *Mesleki Alan İlgi Envanteri'nin Bilgisayar Ortamında Bireye Uyarlanmış Formunun Geliştirilmesi.* (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi, Ankara, Türkiye.

Allison, P. D. (2009). *Missing data. Quantitative methods in psychology*. (Edt: R. E. Millsao & A. Maydeu-Olivares). London: SAGE Publication. pp. 72-89.

Anderson, J. C. & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatoly factor analysis. *Psychometrika, 49*, 155-73.

Arslan, Y. K., Çolak, M., & Bilgin, U. (2022). Determining ability levels by using computerized adaptive testing and paper-and-pencil testing in the distance education: a multi-centred real data study. *Tıp Eğitimi Dünyası, 21*(63), 95-103.

Aybek, E. C. (2016). *Kendini Değerlendirme Envanterinin Bilgisayar Ortamında Bireye Uyarlanmış Test (BOBUT) Olarak Uygulanabilirliğinin Araştırılması.* (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi, Ankara, Türkiye.

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. https://dergipark.org.tr/en/pub/tpdrd/issue/40299/481364

Babcock, B., & Weiss, D. J. (2009). *Termination criteria in computerized adaptive tests: Variable-length CATs are not biased*. In Proceedings of the 2009 GMAC conference on computerized adaptive testing.

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CATs provide efficient and effective measurement. *Journal of Computerized Adaptive Testing, 1*(1), 1-18.

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

Balta, E., & Uçar, A. (2022). Investigation of measurement precision and test length in computerized adaptive testing under different conditions, *E-International Journal of Educational Research, 13*(1), 51-68. https://doi.org/10.19160/e-ijer.1023098

Bedsole, C. B. (2013). *Apples to apples? Comparing the predictive validity of the GMAT and GRE for business schools, and building a better admissions formula*. (Unpublished doctoral dissertation.) University of Alabama, Alabama, United States.

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

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229-255). New York, NY: Routledge.

Bringsjord, E. L. (2001*). Computerized-adaptive versus paper-and-pencil testing environments: An experimental analysis of examinee experience.* (Unpublished doctoral dissertation.) University at Albany, New York, United States.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indexes for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

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.

Büyüköztürk, Ş., Kılıç Çakmak, E., Akgün, Ö. E., Karadeniz, Ş. ve Demirel, F. (2018). *Bilimsel Araştırma Yöntemleri* (24. Baskı). Ankara: Pegem Akademi.

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and recent advances in model selection. *Philosophy of statistics, 7*, 583-605. https://doi.org/10.1016/B978-0-444-51862-0.50018-6.

Chang, H.H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, H.H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229.

Chang, H.H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Measurement in Education, 23*(3), 211-222.

Chang, S. W. & Ansley, T. N. (2006). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement. 40*(1), 71-103.

Chen, J., & Choi, J. (2009). A comparison of maximum likelihood and expected a posteriori estimation for polychoric correlation using Montecarlo simulation. *Journal of Modern Applied Statistical Methods, 8*(1), 337-354.

Chen, L. T., Feng, Y., Wu, P. J., & Peng, C. Y. J. (2020). Dealing with missing data by EM in single-case studies. *Behavior Research Methods, 52*, 131-150.

Cheng, Y., Chang, H.-H., Douglas, J., & Fanmin Guo. (2008). Constraint-weighted a-stratification for computerized adaptive testing with nonstatistical constraints. *Educational and Psychological Measurement, 69*(1), 35–49. doi:10.1177/0013164408322030.

Chew, L. C. (1997). *Validity of computerised adaptive tests for biology achievement testing.* In ERA conference (pp. 24-26). Singapore: Educational Research Association of Singapore (ERAS).

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419–440. doi:10.1177/014662160832780110.1177/0146621608327801.

Choi, S. W., Grady, M. W. & and Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*(1), 37–53.

Cole, D.A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*, 1019-1031.

Costa, D. R., Karino, C. A., Moura, F. A. S., & Andrade, D. F. (2009). *A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing.* In D. J. Weiss (Ed.), Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Crichton, L. I. (1981). *Effect of error in item parameter estimates on adaptive testing* (Unpublished doctoral dissertation), University of Minnesota, Minnesota, United States.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Wadsworth Pub Co.

Çıkrıkçı, N., Yalçın, S., Kalender, İ., Gül, E., Ayan, C., Uyumaz, G., … Kamış, Ö. (2020). Development of a computerized adaptive version of the Turkish driving licence exam. *International Journal of Assessment Tools in Education, 7*(4), 570-587. https://doi.org/10.21449/ijate.716177.

Çoban, E. (2020). *Bilgisayar Temelli Bireyselleştirilmiş Test Yaklaşımlarının Türkiye'deki Merkezi Dil Sınavlarında Uygulanabilirliğinin Araştırılması.* (Yayımlanmamış Doktora Tezi). Ankara Üniversitesi, Ankara, Türkiye.

Dahan Golan, D., Barzillai, M., & Katzir, T. (2018). The effect of presentation mode on children's reading preferences, performance, and self-evaluations. *Computers & Education, 126*, 346–358. https://doi.org/10. 1016/j.compedu.2018.08.001.

Davey, T., & Nering, N. (2002). Controalling item exposure and maintaing item security. In C. N. Mills, M. T. Potenza, J.J. Fremer, & W. C. Ward. (Eds.), *Computer based testing: Building the foundation for future assessments* (pp.165-191). Mahwah, NJ: Lawrence Erlbaum Associates.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items*. (Unpublished doctoral dissertation.) University of Texas, Austin. United States.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning selection of the MCAT. *Applied Psychological Measurement, 27*(5), 335–356.

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

de Beer M. & Visser D. (1998). Comparability of the paper-and-pencil and computerised adaptive versions of the General Scholastic Aptitude Test (GSAT) Senior. *South African Journal of Psychology, 28*(1), 21-27. doi:10.1177/008124639802800104.

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

Demir, E. (2013). Kayıp Verilerin Varlığında Çoktan Seçmeli Testlerde Madde ve Test Parametrelerinin Kestirilmesi: SBS örneği. *Eğitim Bilimleri Araştırmaları Dergisi, 3*(2), 47-68.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*, 1–38.

Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*, 222. https://doi.org/10.1186/2193- 1801-2-222.

Educational Testing Service. (2014). *The research foundation for the GRE revised general test: a compendium of studies.* Author.

Educational Testing Service. (2018). *A snapshot of the individuals who took the GRE revised general test*. Author.

Educational Testing Service. (2019). *GRE guide to the use of scores.* Author.

Educational Testing Service. (2025). *GRE subject tests.* ETS. https://www.ets.org/gre/test-takers/subject-tests/about.html.

Embretson S. E., & Reise S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Earlbaum.

English, R. A., Reckase, M. D., & Patience, W. M. (1977). Application of tailored testing to achievement measurement. *Behavioral Research Methods and Instrumentation, 9*, 158-161.

Ertaş Polat, F. G. (2022). *Çok Aşamalı Bireye Uyarlanmış Testlerde Farklı Koşullardan Elde Edilen Yetenek Kestirimlerinin Karşılaştırılması.* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Finch, H. (2008). Estimation of item response theory parameters in the presence of missing data. *Journal of Educational Measurement, 45* (3), 225-245.

Fotaris, P., Mastoras, T., Mavridis, I. & Manitsaris, A. (2010). *Performance evaluation of the small sample dichotomous IRT analysis in assessment calibration.* 2010 Fifth International Multi-conference on Computing in the Global Information Technology.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning, 37*(5), 1207–1216.

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods, 18*(4), 454-474.

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). Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1647.

Gibbons C., Bower P., Lovell K., Valderas J., & Skevington S. (2016). Electronic quality of life assessment using computer-adaptive testing. *Journal of Medical Internet Research,18*.

GMAC (2023). *The GMAT exam advantage. GMAC.* Erişim adresi: https://www.gmac.com/gmat-other-assessments/about-the-gmat-exam/the-gmat-advantage.

GMAC (2024). *Graduate Management Admission Council. GMAC.* Erişim adresi: https://www.gmac.com/

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347-360.

Gu, L., & Reckase, M. D. (2007). *Designing optimal item pools for computerized adaptive tests with sympson-Hetter exposure control.* Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gür, R., ve Karabay, E. (2018). Motorlu Taşıtlar Sürücü Kursiyerleri Sınavının Simülatif Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulaması. *Journal of Theoretical Educational Science, 11*(2), 201-228. https://doi.org/10.30831/akukeg.395252.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L.(2006*). Multivariate data analysis* (Vol. 6). Upper Saddle River, NJ: PearsonPrentice Hall.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications.* Boston: Kluwer.

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.

Hambleton, R., Swaminathan, R., & Rogers, H.J. (1991). *Fundamentals of ıtem response theory*. California: Sage Pub.

Han, K. (2009). Gradual maximum information ratio approach to item selection in computerized adaptive testing. *Graduate Management Admission Council Research Reports,* RR-09-07.

Han, K. T. & Hambleton, R. K. (2014). *User's manual for wingen 3: windows software that generates IRT model parameters and item responses* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts.

Han, K. T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions, 15*(7), https://doi.org/10.3352/jeehp.2018.15.7.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164. https://doi.org/10.1177/01466216850090020

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement Issuses and Practice, 26(*2), 44-52.

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

Hogan, J., Thornton, N., Diaz-Hoffmann, L., Mohadjer, L., Krenzke, T., Li, J., … Khorramdel, L. (2016). *U.S. program for the ınternational assessment of adult competencies (PIAAC) 2012/2014: main study and national supplement technical report (NCES 2016-036REV).* U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available from http://nces.ed.gov/pubsearch.

Hogan, T. (2014). *Using a computer-adaptive test simulation to ınvestigate test coordinators' perceptions of a high-stakes computer-based testing program*. (Unpublished doctoral dissertation). Georgia State University. Atlanta, United States.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika 30,* 179–185. https://doi.org/10.1007/BF02289447

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International journal of human-computer interaction, 32*(1), 51-62. https://doi.org/10.1080/10447318.2015.1087664

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing, 16*(2), 119-141.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives: structural equation modeling. *A Multidisciplinary Journal, 6*(1), 1-55.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: an introduction to generalized linear models*. Sage Publications

Imai, S., Ito, S., Nakamura, Y., Kikuchi, K., Akagi, Y., Nakasono, H., … Hiramura, T. (2009). *Features of J-CAT (Japanese Computerized Adaptive Test)*. Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved [03.12.2023] from www.psych.umn.edu/psylabs/CATCentral/

Jensen, M. G. (2017). *Extension of the item pocket method allowing for response review and revision to a computerized adaptive test using the generalized partial credit model*. (Unpublished doctoral dissertation). The University of Texas, Austin. United States.

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Educational Sciences: Theory and Practice, 17*(2), 573–596.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability.* (Unpublished doctoral dissertation). Middle East Technical University, Ankara, Türkiye.

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

Kaptan, F. (1993). *Yetenek Kestiriminde Adaptive (Bireyselleştirilmiş) Test Uygulaması ile Kâğıt-Kalem Testi Uygulamasının Karşılaştırılması*. (Yayımlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Karamese, H. (2022). *A comparıson of fınal scorıng methods under the multıstage adaptıve testıng framework*. (Unpublished doctoral dissertation). The University of Iowa, Iowa City, United States.

Karasar, N. (2012). *Bilimsel Araştırma Yöntemi* (23. Baskı). Ankara: Nobel Akademik Yayıncılık.

Kaya, E. (2022). *A comparability and clasiificaiıon analysis of computerized adaptive and conventional paper-based versions of an English language proficiency reading subtest.* (Unpublished doctoral dissertation). İhsan Doğramacı Bilkent University, Ankara, Türkiye.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests*. (Unpublished Doctoral dissertation). The University of Texas, Austin, United States.

Kenny, D. A. & McCoach, D. B. (2003) *Effect of the number of variables on measures of fit in structural equation modeling, structural equation modeling*. 10(3), 333-351, https://doi.org/10.1207/S15328007SEM1003_1.

Keyin, W. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing.* (Unpublished doctoral dissertation). Michigan State University, Michigan, United States.

Kezer, F. (2020). *Yeni Başlayanlar İçin Bilgisayar Ortamında Bireye Uyarlanmış Test Yöntemi* (1. Baskı). Ankara: Pegem Akademi.

Kılıç, A. F., & Uysal, İ. (2019). Comparison of factor retention methods on binary data: A simulation study. *Turkish Journal of Education, 8*(3), 160–179. https://doi.org/10.19128/turje.518636

Kim, J. (2010). *A comparison of computer-based classification testing approaches using mixed- format tests with the generalized partial credit model*. (Unpublished doctoral dissertation). The University of Texas, Austin, United States.

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.

Kim, S., & Moses, T. (2014). An investigation of the impact of misrouting under two-stage multistage testing: A simulation study: An investigation of the impact of misrouting. *ETS Research Report Series, 2014*(1), 1–13.

Kingsbury, G. G., & Houser, R. L. (2005). Assessing the utility of ıtem response models: computerized adaptive testing. educational measurement: *Issues and Practice, 12*(1), 21–27. doi:10.1111/j.1745-3992.1993.tb00520.x

Kingsbury, G., Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-75.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices.* New York: Springer Science and Business Media.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. *Organizational Research Methods, 9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Lau, C. A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model*. In Annual Meeting of the American Educational Research Association. Canada.

Lee, C., & Han, K. C. T. (2024). Evaluating effectiveness of standard error of score estimation as a CAT termination criterion. *Journal of Computerized Adaptive Testing, 11*(2), 13-29.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 37*(4).

Linacre, J. M. (2004). Rasch model estimation: further topics. *Journal of Applied Measurement, 5* (1), 95-110.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association, 83*(404), 1198-1202.

Little, R. J. ve Rubin, D. B. (1987). *Statistical analysis with missing data*. (2nd ed). New York: Wiley

Liu, H.-Y., You, X.-F., Wang, W.-Y., Ding, S.-L., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*(2), 152–172.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Lord, F. M. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika, 48,* 477-482.

Lord, F.M. (1986). Maximum likelihood and bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Luecht, R. M. & Sireci, S. G. (2011). *A review of models for computer-based testing*. (Research Report RR-2011-12). New York: The College Board.

Lunz, M. E., Bergstrom, B. A., & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement,16*(1), 33–40.

Mair, P. (2018). *Modern psychometrics with R.* Springer International Publishing AG.

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185–187.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and Introduction. *Applied Psychological Measurement, 23*(3), 187–194. doi:10.1177/01466219922031310.

Mertens, D. M. (2010). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods.* (3rd ed.). SAGE Publications.

Mislevy, R. J. & Wu, P. K. (1996). *Missing responses and IRT ability estimation: omits, choice, time limits, adaptive testing.* ETS Research Report RR-96-30-ONR, Princeton, NJ: Educational Testing Service.

Molina, M. T. L-M. (2009). A computer-adaptive vocabulary test. *Indian Journal of Applied Linguistics, 35*(1), 121-138.

Mullis, I. V. S., Martin, M. O., Goh, S., & Prendergast, C. (2017). *PIRLS 2016 encyclopedia: Education policy and curriculum in reading.* Boston College. TIMSS & PIRLS International Study Center.

Ogunjimi, M. O., Ayanwale, M. A., Oladele, J. I., Daramola, D. S., Jimoh, I. M. & Owolabi, H. O. (2021). Simulated evidence of computer adaptive test length: Implications for high stakes assessment in Nigeria. *Journal of Higher Education Theory and Practice, 21*(2). 202-212.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Oswald, F. L., Shaw, A., & Farmer, W. L. (2014). Comparing simple scoring with IRT scoring of personality measures. *Applied Psychological Measurement, 39*(2), 144-154.

Öcal, İ. Ü., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications. *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 395-407.

ÖSYM (2018a). 2017 ALES Sonbahar Değerlendirme Raporu. Erişim adresi: https://dokuman.osym.gov.tr/pdfdokuman/2018/GENEL/ALESRaporu_13022018.pdf

ÖSYM (2018b). 2018 ALES/2 Değerlendirme Raporu. Erişim adresi: https://dokuman.osym.gov.tr/pdfdokuman/2018/GENEL/ales2degrapor24122018.pdf

ÖSYM (2023). ÖSYM Tanıtım Bülteni. Erişim adresi: https://www.osym.gov.tr/TR,25533/tanitim-bulteni.html. Erişim Tarihi 21.09.2023

Özbaşı, D. (2014). *Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğine İlişkin Bir Araştırma*. (Yayınlanmamış Doktora tezi). Ankara Üniversitesi, Ankara, Türkiye.

Özbaşı, D., ve Demirtaşlı, N. (2015). Bilgisayar Okuryazarlığı Testinin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Geliştirilmesi. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2). https://doi.org/10.21031/epod.79491.

Özdemir, B. (2016). *Comparison of different unidimensional-CAT algorithms measuring students' language abilities: Post-hoc simulation study.* European Proceedings of Social and Behavioural Sciences.

Pallant, J. (2020). *SPSS survival manual. A step by step guide to data analysis using SPSS*. Routledge. https://doi.org/10.4324/9781003117452

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing.* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, United States.

Petersen, M. A., Gamper, E. M., Costantini, A., Giesinger, J. M., Holzner, B., Johnson, C., ... & EORTC Quality of Life Group. (2016). An emotional functioning item bank of 24 items for computerized adaptive testing (CAT) was established. *Journal of Clinical Epidemiology, 70*, 90-100.

Pramono, A. (2023). Computerızed adaptive test (CAT) optimization using the efficiency balanced information (EBI) method. *Proceeding of İnternational Conference on Education, Socety And Humanity, 1*(1), 320-330.

Puhan, G., Boughton, K. A., & Kim, S. (2005). *Evaluating the comparability of paper-and-pencil and computerized versions of a large-scale certification test*. ETS Research Report Series, 2005(2), i-15.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications, *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.2307/1164671.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests.* Paper presented at the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D. (2009). *Statistics for social and behavioral sciences: Multidimensional item response theory*. Springer-Verlag New York.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127-141.

Reeve, B., & Fayers, P. (2005). Applying item response theory modelling for evaluating questionnaire item and scale properties. In P. M. Fayers & R. D. Hays (Eds.), *Assessing quality of life in clinical trials: Methods and practice* (2nd ed., pp. 55–73). Oxford, UK: Oxford University Press.

Reise, S. P., & Revicki, D. A. (2015). *Handbook of item response theory modeling. applications to typical performance assessment.* New York, NY: Routledge.

Revicki, D. A., Chen, W.-H., & Tucker, C. (2014). Developing item banks for patient-reported health outcomes. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 334–363). New York, NY: Routledge.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4) 311–327.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.

Rotou, O., Patsula, L., Steffen, M. & Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests.* ETS, Research Report.

Rubin, D. B. (1987). *Multiple ımputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc.

Rubinstein, R. Y., & Kroese, D. P. (2007). *Simulation and the Monte Carlo method*. Wiley.

Rudner, L. M. (2009). Implementing the graduate management admission test computerized adaptive test. *Elements of Adaptive Testing*, 151–165.

Rudner, L. M. (2012). *Demystifying the GMAT: computer-based testing terms*. Graduate Management Admission Council.

Sari, H. İ., & Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. Educational Sciences: *Theory & Practice, 17*(5). https://doi.org/10.12738/estp.2017.5.0484

Sayman Ayhan, A. (2015). *Comparability of scores from cat and paper and pencil implementations of student selection examination to higher education*. (Unpublished doctoral dissertation). İhsan Doğramacı Bilkent University, Ankara, Türkiye.

Schaffer, J. L. (1997). *Analysis of ıncomplete multivariate data*. London: Chapman & Hall.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online, 8*(2), 23-74.

Scullard, M. G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory*. (Unpublished doctoral dissertation). University of Minnesota, Minnesota. United States.

Segall, D. O. (2005). Computerized adaptive testing. *Encyclopedia of Social Measurement, 1*, 429–438.

Seo, D. G., & Choi, J. (2018). Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination. *Journal of Educational Evaluation for Health Professions, 15*(14). https://doi.org/10.3352/jeehp.2018.15.14

Shin, C. D., Chien, Y., & Way, W. D. (2012). *A comparison of three content balancing methods for fixed and variable length computerized adaptive tests.* Paper presented at the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Shin, C. D., Chien, Y., Way, W. D., & Swanson, L. (2009). *Weighted penalty model for content balancing in CATs.* Retrieved from http://education.pearson assessments.com/NR/rdonlyres/99A4327B-5968-4AB2-A8CD-8D502D22C2DE/0/ WeightedPenaltyModel.pdf.

Slater, S. C. (2001). *Pretest item calibration within the computerized adaptive testing environment.* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, United States.

Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy, 43*, 679–685. https://doi.org/10.1016/j. beth.2011.10.001

Smits N, Cuijpers P, & van Straten A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research, 188*(1). 147-155.

Springer, T., & Urban, K. (2014). Comparison of the EM algorithm and alternatives. *Numerical Algorithm 67*(2), 335–364. https://doi.org/ 10.1007/s11075-013-9794-8

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained ıtem selection in adaptive testing. *Applied Psychological Measurement, 17*, 277-292.

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education, 151*, 103861. https://doi.org/10.1016/j.compedu.2020.103861.

Sulak, S. (2013). *Bireyselleştirilmiş Bilgisayarlı Test Uygulamalarında Kullanılan Madde Seçme Yöntemlerinin Karşılaştırılması*. (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Şahin, A. & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory and Practice, 15*(6), 1585-1595.

Şen, Ö. F. (2022). *Bilgisayarda Bireyselleştirilmiş Test Uygulamalarında Maddeyi Yeniden Cevaplayabilmenin Ölçme Hatasına ve Yanlılığına Etkisinin İncelenmesi*. (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Şenel, S. (2017). *Bilgisayar Ortamında Bireye Uyarlanmış Testlerin Görme Engelli Öğrencilere Uygunluğunun İncelenmesi*. (Yayınlanmamış Doktora Tezi). Ankara Üniversitesi, Ankara, Türkiye.

Şimşek, A. S. (2017). *Becerilere Güven Mesleki İlgi Envanterinin Uyarlanması ve Bilgisayarlı Bireyselleştirilmiş Test Uygulamasının Geliştirilmesi*. (Yayınlanmamış Doktora Tezi). Ankara Üniversitesi, Ankara, Türkiye.

Taylor, C., Jamieson, J., Eignor, D. R., & Kirsch, I. (1998). *The relationship between computerfamiliarity and performance on computer-based TOEFL test tasks* (ETS RR-98-08). Princeton, NJ: ETS.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. Dorans, D. Eignor, R. Flaugher, B. Green, R. Mislevy, L. Steinberg, & D. Thissen (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–133). Routledge.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (Chap. 3, pp. 73–140). Lawrence Erlbaum Associates,Inc., Publishers, New Jersey.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*(1). Retrieved from: http://pareonline.net/getvn.asp?v=16&n=1.

Thompson, S., Thurlow, M., & Moore, M. (2003). *Using computer-based tests with students with disabilities.* NCEO Policy Directions.

Threlfall, J., Pool, P., Homer, M., & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics, 66*(3), 335-348.1-9.

Tsaousis, I., Sideridis, G. D., & AlGhamdi, H. M. (2021). Evaluating a computerized adaptive testing version of a cognitive ability test using a simulation study. *Journal of Psychoeducational Assessment, 39*(8), 954-968.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

Uttaro, T. & Lehman, A. (1999). Graded response modeling of the quality of life interview. *Evaluation and Program Planning, 22*(1999), 41-52.

Valdivia Medinaceli, M. B. (2023). *Addressıng current methodologıcal challenges ın ılsa's transıtıon to adaptıve testıng*. (Unpublished doctoral dissertation). Indiana University. United States.

van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory.* Springer Science Business Media. New York.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201–216. doi:10.1007/bf02294775.

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.

van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing.* (Vol 10, pp. 978-0). New York: Springer.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. computerized adaptive testing: *Theory and Practice*, 1–25. doi:10.1007/0-306-47531-6_1.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing, *Journal of Educational and Behavioral Statistics*, *22*(2), 203–226.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? Measurement Interdisciplinary *Research and Perspectives 7*(2) https://doi.org/10.1080/15366360903117079.

Wagner, I., Loesche, P., & Bißantz, S. (2022). Low-stakes performance testing in Germany by the VERA assessment: Analysis of the mode effects between computer-based testing and paper-pencil testing. *European Journal of Psychology of Education, 37*(2), 531-549.

Wainer, H. (Ed). (2000). *Computerized adaptive testing: A primer.* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Wang, C., Chang, H.-H., & Boughton, K. A. (2010). Kullback–Leibler information and its applications in multi-dimensional adaptive testing. *Psychometrika, 76*(1), 13–39. doi:10.1007/s11336-010-9186-0.

Wang, H. B., Kuo, B. C., Tsai, Y., & Liao, C. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *The Turkish Online Journal of Educational Technology, 11*(4). 1-13.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing*. (Unpublished doctoral dissertation). Michigan State University, Michigan, United States.

Wang, S., & Zhang, L. (2017). *Effects of ignoring discrimination parameter in cat item selection on student scores*. National Council on Measurement in Education (NCME) conferences, San Antonio, TX.

Wang, T. (1997). *Essentially unbiased EAP estimates in computerized adaptive testing*. http://iacat.org/sites/default/files/biblio/wa97-01.pdf

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods computerized adaptive testing. *Journal of Educational Measurement, 35*(2), 109-135.

Wang, T., Hanson, B. A., & Lau, C.-M. A. (1999). Reducing bias in cat trait estimation: a comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278. https://doi.org/10.1177/01466219922031383

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels formultidimensional assessment*. (Unpublished doctoral dissertation). The University of North Carolina Greensboro, North Carolina, United States.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educatıonal Measurement, 21*(4), 361-375.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Weissman, A. (2006). A feedback control strategy for enhancing ıtem selection efficiency in computerized adaptive testing. *Applied Psychological Measurement, 30*(2), 84–99. doi:10.1177/0146621605282774.

Wise, S. L., Kingsbury, G. G. & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practice, 34*(4), 41-48. https://doi.org/10.1111/emip.12094

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806–838. https://doi.org/10.1177/0011000006288127.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

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

Yasuda, J. I., & Hull, M. M. (2021). *Balancing content of computerized adaptive testing for the Force Concept Inventory*. In Proceedings of the 2021 Physics Education Research Conference. AIP.

Yen, Y. C., Ho, R. G., Laio, W. W., Chen, L. J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement, 36*(2), 75-87.

Yi, Q., Wang, T., & Ban, J. C. (2000). *Effects of scale transformation and test termination rule on the precision of ability estimates in CAT*. ACT Research Report Series.

Yiğiter, M. S. (2023). *Sabit ve Anında Bireyselleştirilmiş Çok Aşamalı Testlerin Karşılaştırılması.* (Yayınlanmamış Doktora Tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Yu, C.Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. (Unpublished doctoral dissertation). University of California, Los Angeles, United States.

Zhang, J. (2013). A sequential procedure for detecting compromised items in the item pool of a cat system. *Applied Psychological Measurement, 38*(2), 87–104. doi:10.1177/0146621613510062.

Zheng Y, Chang C-H & Chang H-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Qual Life Res, 22*(3), 491-499.

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* (Research Report 2012–6). Iowa City: ACT, Inc.

Žitný, P., Halama, P., Jelínek, M., & Květon, P. (2012). Validity of cognitive ability tests–comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. *Studia Psychologica, 54*(3), 181-194.

EKLER

EK 1. Etik Kurul İzni

ANKARA ÜNİVERSİTESİ
ALT ETİK KURULU
KARAR ÖRNEĞİ

Karar Tarihi : **06.05.2024**
Toplantı Sayısı : **14**
Karar Sayısı : **132**

132- Üniversitemiz Eğitim Bilimleri Fakültesi, öğretim üyelerinden **Doç. Dr. Seher YALÇIN**'ın danışmanlığını yaptığı, doktora öğrencisi **Tansu ALAN**'ın, "Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavının Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi" doktora tezi ile ilgili 22/03/2024 tarihli "İnsan Üzerinde Yapılan Klinik Dışı Araştırmalar Başvuru Formu" Etik Kurulumuzca incelendi.

Üniversitemiz Eğitim Bilimleri Fakültesi, öğretim üyelerinden **Doç. Dr. Seher YALÇIN**'ın danışmanlığını yaptığı, doktora öğrencisi **Tansu ALAN**'ın, "Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavının Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi" doktora tezi ile ilgili araştırma protokolüne uyulması ve etik onay tarihinden itibaren geçerli olması koşuluyla uygulanmasının etik açıdan uygun olduğuna oy birliği ile karar verilmiştir.

ASLININ AYNIDIR
06/05/2024

Prof. Dr. Muharrem ÖZEN
Ankara Üniversitesi
Etik Kurulu Başkanı

EK 2. ÖSYM Veri Kullanım İzni

T.C.
ÖLÇME, SEÇME VE YERLEŞTİRME MERKEZİ BAŞKANLIĞI
Araştırma Geliştirme Daire Başkanlığı

Sayı : E-92968509-605-353259
Konu : Tansu ALAN - Veri Talebi

ANKARA ÜNİVERSİTESİ REKTÖRLÜĞÜNE
Dögol Caddesi 06100
ÇANKAYA/ANKARA

Üniversiteniz Eğitim Bilimleri Enstitüsü Eğitim Bilimleri Anabilim Dalı Eğitimde Ölçme ve Değerlendirme doktora programı öğrencisi Tansu ALAN'ın "Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavının Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi" konulu tez çalışması kapsamında Başkanlığımızdan talep etmiş olduğu 2021 yılına ait ALES/1, ALES/2, ALES/3 sınav verilerinin bahsi geçen öğrenci ile paylaşılmasına ve bu verilerin ilgili tezde kullanılmasına ilişkin gereken iznin verilmesi hususu "Veri Paylaşımı Değerlendirme Komisyonu"nca uygun görülmüştür.

Bilgilerinizi rica ederim.

Prof. Dr. Mustafa DOĞAN
Başkan a.
Başkan Yardımcısı

Bu belge, güvenli elektronik imza ile imzalanmıştır.

Belge Doğrulama Kodu : BSRBS7JKYS Pin Kodu : 09603
Belge Takip Adresi : http://evrakdogrulama.osym.gov.tr/envision.sorgula/belgedogrulama.aspx?

Adres : Üniversiteler Mahallesi İhsan Doğramacı Bulvarı No:4D Bilkent / ANKARA
Faks : 0312 2988810
Web : www.osym.gov.tr
Kep Adresi : osym@hs01.kep.tr

EK 3. ALES 2021 Dönem 1 Kayıp Veri Oranları

ALES 2021 Dönem 1 Alt Testlere Ait Kayıp Veri Oranları

	Sayısal				Sözel	
Madde	n	Kayıp veri sayısı	Kayıp veri oranı (%)	N	Kayıp veri sayısı	Kayıp veri oranı (%)
Y1	3500	494	14.1	Y51	138	3.9
Y2	3500	786	22.5	Y52	160	4.6
Y3	3500	1339	38.3	Y53	458	13.1
Y4	3500	730	20.9	Y54	352	10.1
Y5	3500	745	21.3	Y55	171	4.9
Y6	3500	974	27.8	Y56	216	6.2
Y7	3500	873	24.9	Y57	201	5.7
Y8	3500	1124	32.1	Y58	193	5.5
Y9	3500	2333	66.7	Y59	342	9.8
Y10	3500	1571	44.9	Y60	522	14.9
Y11	3500	1093	31.2	Y61	383	10.9
Y12	3500	1534	43.8	Y62	482	13.8
Y13	3500	1605	45.9	Y63	523	14.9
Y14	3500	1657	47.3	Y64	638	18.2
Y15	3500	1140	32.6	Y65	989	28.3
Y16	3500	1604	45.8	Y66	480	13.7
Y17	3500	861	24.6	Y67	355	10.1
Y18	3500	1771	50.6	Y68	407	11.6
Y19	3500	2173	32.1	Y69	372	10.6
Y20	3500	1903	54.4	Y70	319	9.1
Y21	3500	1204	34.4	Y71	379	10.8
Y22	3500	1691	48.3	Y72	604	17.3
Y23	3500	2045	58.4	Y73	765	21.9
Y24	3500	2242	64.1	Y74	469	13.4
Y25	3500	1389	39.7	Y75	529	15.1
Y26	3500	1255	35.9	Y76	692	19.8
Y27	3500	1630	46.6	Y77	477	13.6
Y28	3500	1968	56.2	Y78	437	12.5
Y29	3500	1968	56.2	Y79	672	19.2
Y30	3500	2153	61.5	Y80	686	19.6
Y31	3500	2009	57.4	Y81	886	25.3
Y32	3500	2176	62.2	Y82	925	26.4
Y33	3500	2564	73.3	Y83	996	28.5
Y34	3500	2327	66.5	Y84	887	25.3
Y35	3500	2234	63.8	Y85	908	25.9
Y36	3500	2265	64.7	Y86	913	26.1
Y37	3500	2404	68.7	Y87	1041	29.7
Y38	3500	1839	52.5	Y88	956	27.3
Y39	3500	1924	55.0	Y89	1153	32.9
Y40	3500	2119	62.6	Y90	1200	34.3
Y41	3500	1630	46.6	Y91	1176	33.6
Y42	3500	2048	58.5	Y92	1200	34.3
Y43	3500	1913	54.7	Y93	1703	48.7
Y44	3500	2070	59.1	Y94	1713	48.9
Y45	3500	2270	64.9	Y95	1638	46.8
Y46	3500	2261	64.6	Y96	1726	49.3
Y47	3500	2420	69.1	Y97	1506	43.0
Y48	3500	2566	73.3	Y98	1560	44.6
Y49	3500	2005	57.3	Y99	1726	49.3
Y50	3500	1633	46.7	Y100	1556	44.5

EK 4. ALES 2021 Dönem 2 Kayıp Veri Oranları

ALES 2021 Dönem 2 Alt Testlere Ait Kayıp Veri Oranları

	Sayısal				Sözel	
Madde	n	Kayıp veri sayısı	Kayıp veri oranı (%)	N	Kayıp veri sayısı	Kayıp veri oranı (%)
Y1	3500	458	13.1	Y51	151	4.3
Y2	3500	849	24.3	Y52	188	5.4
Y3	3500	1347	38.5	Y53	386	11.0
Y4	3500	825	23.6	Y54	765	21.9
Y5	3500	1323	37.8	Y55	397	11.3
Y6	3500	1047	29.9	Y56	240	6.9
Y7	3500	958	27.4	Y57	244	7.0
Y8	3500	1730	49.4	Y58	232	6.6
Y9	3500	1503	42.9	Y59	439	12.5
Y10	3500	1681	48.0	Y60	403	11.5
Y11	3500	2195	62.7	Y61	906	25.9
Y12	3500	2436	69.6	Y62	515	14.7
Y13	3500	1745	49.9	Y63	583	16.7
Y14	3500	1269	36.3	Y64	665	19.0
Y15	3500	1286	36.7	Y65	526	15.0
Y16	3500	1766	50.5	Y66	359	10.3
Y17	3500	1400	40.0	Y67	513	14.7
Y18	3500	1738	49.7	Y68	479	13.7
Y19	3500	1409	40.3	Y69	489	14.0
Y20	3500	2677	76.5	Y70	511	14.6
Y21	3500	1359	38.8	Y71	295	8.4
Y22	3500	2390	68.3	Y72	586	16.7
Y23	3500	2108	60.2	Y73	714	20.4
Y24	3500	1975	56.4	Y74	730	20.9
Y25	3500	2573	73.5	Y75	791	22.6
Y26	3500	2651	75.7	Y76	671	19.2
Y27	3500	1328	37.9	Y77	704	20.1
Y28	3500	1296	37.0	Y78	684	19.5
Y29	3500	2331	66.6	Y79	893	25.5
Y30	3500	2947	84.2	Y80	863	24.7
Y31	3500	2627	75.1	Y81	875	25.0
Y32	3500	2328	66.5	Y82	798	22.8
Y33	3500	2318	66.2	Y83	1056	30.2
Y34	3500	2529	72.3	Y84	1053	30.1
Y35	3500	1483	42.4	Y85	1018	29.1
Y36	3500	1701	48.6	Y86	1020	29.1
Y37	3500	1555	44.4	Y87	1050	30.0
Y38	3500	1880	53.7	Y88	1010	28.9
Y39	3500	2000	57.1	Y89	1112	31.8
Y40	3500	2381	68.0	Y90	1109	31.7
Y41	3500	1842	52.6	Y91	1097	31.3
Y42	3500	1647	47.1	Y92	1121	32.0
Y43	3500	1777	50.8	Y93	1695	48.4
Y44	3500	1448	41.4	Y94	1581	45.2
Y45	3500	2133	60.9	Y95	1525	43.6
Y46	3500	2272	64.9	Y96	1680	48.0
Y47	3500	2275	65.0	Y97	2306	65.9
Y48	3500	2056	58.7	Y98	2378	67.9
Y49	3500	1954	55.8	Y99	2383	68.1
Y50	3500	2015	57.6	Y100	2377	67.9

EK 5. ALES 2021 Dönem 3 Kayıp Veri Oranları

ALES 2021 Dönem 3 Alt Testlere Ait Kayıp Veri Oranları

	Sayısal				Sözel	
Madde	n	Kayıp veri sayısı	Kayıp veri oranı (%)	n	Kayıp veri sayısı	Kayıp veri oranı (%)
Y1	3000	333	11.1	Y51	132	4.4
Y2	3000	884	29.5	Y52	109	3.6
Y3	3000	686	22.9	Y53	149	5.0
Y4	3000	562	18.7	Y54	188	6.3
Y5	3000	907	30.2	Y55	142	4.7
Y6	3000	1392	46.4	Y56	164	5.5
Y7	3000	659	22.0	Y57	198	6.6
Y8	3000	754	25.1	Y58	155	5.2
Y9	3000	996	33.2	Y59	251	8.4
Y10	3000	1300	43.3	Y60	539	18.0
Y11	3000	1167	38.9	Y61	487	16.2
Y12	3000	1011	33.7	Y62	362	12.1
Y13	3000	1566	52.2	Y63	394	13.1
Y14	3000	1545	51.5	Y64	542	18.1
Y15	3000	1149	38.3	Y65	376	12.5
Y16	3000	1678	55.9	Y66	288	9.6
Y17	3000	1449	48.3	Y67	433	14.4
Y18	3000	1189	39.6	Y68	360	12.0
Y19	3000	1602	53.4	Y69	341	11.4
Y20	3000	1087	36.2	Y70	220	7.3
Y21	3000	1343	44.8	Y71	329	11.0
Y22	3000	1665	55.5	Y72	422	14.1
Y23	3000	1010	33.7	Y73	329	11.0
Y24	3000	1390	46.3	Y74	396	13.2
Y25	3000	1151	38.4	Y75	407	13.6
Y26	3000	1166	38.9	Y76	339	11.3
Y27	3000	852	28.4	Y77	570	19.0
Y28	3000	1021	34.0	Y78	404	13.5
Y29	3000	1406	46.9	Y79	647	21.6
Y30	3000	1624	54.1	Y80	617	20.6
Y31	3000	1478	49.3	Y81	710	23.7
Y32	3000	1435	47.8	Y82	627	20.9
Y33	3000	1548	51.6	Y83	642	21.4
Y34	3000	1524	50.8	Y84	660	22.0
Y35	3000	1723	57.4	Y85	637	21.2
Y36	3000	1763	58.8	Y86	839	28.0
Y37	3000	2245	74.8	Y87	783	26.1
Y38	3000	1598	53.3	Y88	855	28.5
Y39	3000	1584	52.8	Y89	1045	34.8
Y40	3000	1555	51.8	Y90	1083	36.1
Y41	3000	1372	45.7	Y91	1121	37.4
Y42	3000	1516	50.5	Y92	1031	34.4
Y43	3000	1389	46.3	Y93	1221	40.7
Y44	3000	1542	51.4	Y94	1255	41.8
Y45	3000	1722	57.4	Y95	1381	46.0
Y46	3000	1919	64.0	Y96	1434	47.8
Y47	3000	1880	62.7	Y97	1171	39.0
Y48	3000	2066	68.9	Y98	1232	41.1
Y49	3000	1544	51.5	Y99	1163	38.8
Y50	3000	1380	46.0	Y100	1255	41.8

BENZERLİK BİLDİRİMİ

“Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavının Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliğinin İncelenmesi" başlıklı tezimin ana bölümü (ön bölüm, kaynaklar ve ekler hariç) Turnitin İntihali Engelleme Programı aracılığıyla incelenmiş ve ilgili rapor danışmanım tarafından da kontrol edilmiştir. Kontrol sırasında (1) “Beş sözcükten daha az olan benzeşmeler” (2) “Kaynaklar” (3) “Doğrudan Alıntılar” dışarıda tutulmuştur. Benzerlik kontrolüne ilişkin rapordan elde edilen sonuçlar aşağıda sunulmuştur.

Rapor Tarihi	**01.12.2025**
Gönderim Numarası	**2696354406**
Sayfa Sayısı	**130**
Sözcük Sayısı	**39075**
Karakter Sayısı	**335221**
Benzerlik Oranı	**%9**
Savunma Tarihi	**08.01.2026**

Yukarıda belirtilen sonuçları gösteren Turnitin İntihal Tespit Programı’na ilişkin orijinal raporu, sonuçlarda herhangi bir değişiklik yapmaksızın bu beyanım ekinde Enstitüye teslim ettiğimi, tezimin %10’dan fazla benzerlik oranı içerdiğinin belirlenmesi durumunda, bundan doğabilecek tüm yasal sorumluluğu kabul ettiğimi bildirir, saygılarımı sunarım.

Öğrencinin Adı Soyadı: Tansu ALAN

Tarih: 01.12.2026

İmza:

ÖZGEÇMİŞ

Kişisel Bilgiler

Adı ve Soyadı: Tansu ALAN

Yabancı Dil: YÖKDİL İngilizce, 80

İletişim:

İş Deneyimi

Çalıştığı Kurum: Adıyaman Üniversitesi / Rektörlük: Toplam Kalite Yönetimi Koordinatörlüğü, Büyük Veri Yönetimi Ofisi

Unvan: Öğretim Görevlisi

Akademik Bilgiler

Öğrenim Durumu:

Derece	Bölüm/Program	Üniversite	Yıl
Doktora	Eğitimde Ölçme ve Değerlendirme	Ankara Üniversitesi	2021-2026
Yüksek Lisans	Eğitimde Ölçme ve Değerlendirme	Hasan Kalyoncu Üniversitesi	2018-2021
Lisans	İlköğretim Matematik Öğretmenliği	Adıyaman Üniversitesi	2014-2018

Yayınlar:

Makale

1. **Alan, T.**, & Yalçın, S. (2025). Lisansüstü Programlara Başvuruda Kullanılan Sınavların Karşılaştırmalı İncelenmesi: GRE, GMAT ve ALES Örneği. *Milli Eğitim Dergisi*, *54*(247), 1801-1832. https://doi.org/10.37669/milliegitim.1577420

2. **Alan, T.** (2025). A Comparative Analysis of Different Machine Learning Models for Classifying Student Achievement. *Adıyaman University Journal of Educational Sciences*, *15*(1), 159-184. https://doi.org/10.17984/adyuebd.1551029

3. Kiye, S., Zorbaz, S. D., & **Alan, T.** (2025). Çocuklar için Psikolojik İyi Oluş Ölçeği: Bir Ölçek Uyarlama Çalışması. *Batı Anadolu Eğitim Bilimleri Dergisi*, *16*(1), 1774-1794. https://doi.org/10.51460/baebd.1577138

4. **Alan, T.**, & Akbaş, U. (2025). Examination of Misconceptions in the Field of Alternative Measurement and Evaluation with A Four-Tier Test. *Kalem Eğitim ve İnsan Bilimleri Dergisi,15*(1), 187-204.DOI: 10.23863/kalem.2024.294

Bildiri

1. **Alan T.,** ve Yalçın, S. (2024-Özet Bildiri). Lisansüstü Programlara Başvuruda Kullanılan Sınavların Karşılaştırmalı İncelenmesi. Uluslararası Ölçme, Seçme ve Yerleştirme Sempozyumu, Ankara, Türkiye.

2. **Alan, T.** (2024- Özet Bildiri). Öğrenci Başarısının Sınıflandırılmasında Farklı Makine Öğrenmesi Yöntemlerinin Karşılaştırılması. IX. Uluslararası Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Eskişehir, Türkiye.

3. Keskinkılıç M. E., **Alan T.,** Demirtaş Zorbaz S., Akın Arıkan Ç. ve Şengül Avşar A. (2024 -Tam Metin Bildiri). Predictors of Gaming Disorder Among University Students. 9th International Conference on Social Sciences, Humanities and Education, Vienna, Austria.

4. **Alan, T.** (2023- Tam Metin Bildiri). Teachers' Opinions On The Education Of Syrian Refugee Students During The Pandemic, Uluslararası Eğitim Yönetimi Forumu Kongresi (EYFOR-14), 03-07 Mayıs 2023, Antalya, Türkiye

5. Turan, E. D. ve **Alan, T.** (2023- Tam Metin Bildiri). Erken Çocukluk Döneminde Ebeveynleri Tarafından Hikâye Okunan Çocukların Davranışlarında ve Gelişim Alanlarında Meydana Gelen Değişimler Hakkında Ebeveynleri ile Görüşülmesi, Uluslararası Eğitim Yönetimi Forumu Kongresi (EYFOR-14), 03-07 Mayıs 2023, Antalya, Türkiye

6. Taşkıran, E., **Alan, T**., Atakan, H., ve Baglama, Y. (2019- Özet Bildiri). Suriyeli Öğrencilerin Dil Farklılıklarının Matematik Başarısına Etkisi, 6. Uluslararası Multidisipliner Çalışmaları Kongresi, 26-27 Nisan 2019, Hasan Kalyoncu Üniversitesi, Gaziantep, Türkiye

Proje

TÜBİTAK 1001 Projesi (Nisan 2023- Mart 2024- **Bursiyer**): Melek Mi Şeytan Mı? Çevrimiçi Rekabetçi Oyunların Üniversite Öğrencilerinin Sosyal- Duygusal Gelişimine Etkisi (Yürütücü: Doç. Dr. Selen Demirtaş Zorbaz).

Yüksek Lisans Tezi

Alan, T. (2021). Tamamlayıcı Ölçme ve Değerlendirme Alanındaki Kavram Yanılgılarının Dört Aşamalı Testle İncelenmesi. (Yayınlanmamış yüksek lisans tezi). Hasan Kalyoncu Üniversitesi, Gaziantep.