HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

DEĞİŞEN MADDE FONKSİYONU GÖSTEREN MADDE ORANININ BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI VE ÇOK AŞAMALI TESTLER ÜZERİNDEKİ ETKİSİ

Başak ERDEM KARA

Doktora Tezi

Ankara, 2019

Liderlik, araştırma, inovasyon, kaliteli eğitim ve değişim ile

Daha ileriye... En İyiye...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Eğitim Bilimleri Ana Bilim Dalı
Eğitimde Ölçme ve Değerlendirme Programı

DEĞİŞEN MADDE FONKSİYONU GÖSTEREN MADDE ORANININ BİREYSELLEŞTİRİLMİŞ BİLGİSAYARLI VE ÇOK AŞAMALI TESTLER ÜZERİNDEKİ ETKİSİ

THE EFFECT OF ITEM RATIO INDICATING DIFFERENTIAL ITEM FUNCTIONING ON COMPUTER ADAPTIVE AND MULTI STAGE TESTS

Başak ERDEM KARA

Doktora Tezi

Ankara, 2019

Kabul ve Onay

Eğitim Bilimleri Enstitüsü Müdürlüğüne,

Başak ERDEM KARA'nın hazırladığı "Değişen Madde Fonksiyonu Gösteren Madde Oranının Bireyselleştirilmiş Bilgisayarlı ve Çok Aşamalı Testler Üzerindeki Etkisi" başlıklı bu çalışma jürimiz tarafından **Eğitim Bilimleri Ana Bilim Dalı, Eğitimde Ölçme ve Değerlendirme Bilim Dalında Doktora Tezi** olarak kabul edilmiştir.

Jüri Başkanı	Prof. Dr. Selahattin GELBAL	
Jüri Üyesi (Danışman)	Prof. Dr. Nuri DOĞAN	
Jüri Üyesi	Prof. Dr. Hülya KELECİOĞLU	
Jüri Üyesi	Prof. Dr. Şeref TAN	
Jüri Üyesi	Doç. Dr. Deniz GÜLLEROĞLU	

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim, Öğretim ve Sınav Yönetmeliği'nin ilgili maddeleri uyarınca yukarıdaki jüri üyeleri tarafından 04 / 10 / 2019 tarihinde uygun görülmüş ve Enstitü Yönetim Kurulunca ..... / ..... / ..... tarihi itibarıyla kabul edilmiştir.

Prof. Dr. Ali Ekber ŞAHİN
Eğitim Bilimleri Enstitüsü Müdürü

Öz

Bu çalışmanın amacı, değişen madde fonksiyonu (DMF) gösteren maddelerin bulunduğu testlerde bireyselleştirilmiş bilgisayarlı test yaklaşımlarının (BBT ve ÇAT) performansının farklı koşullar altında incelenmesidir. Hem bireyselleştirilmiş bilgisayarlı testlerin (BBT) hem de çok aşamalı testlerin (ÇAT) DMF'li maddelerden ne şekilde etkilendiği farklı test uzunlukları (10-20-30 ve 40 madde) ve DMF'li madde oranları (%10 - %20 ve %30) altında araştırılmıştır. Bu amaçla, simülasyon yöntemi ile 5000 kişilik birey grubuna ilişkin yetenek parametreleri ve 600 maddelik bir havuz oluşturulmuş, havuzdaki maddelerin bir kısmı araştırmacı tarafından güçlük parametreleri değiştirilerek DMF'li hale getirilmiştir. Daha sonra, bir BBT ve iki ÇAT ortamı (1-3-3 ve 1-2-4 panel desenleri) RStudio üzerinde 'xxIRT', 'catR' ve 'mstR' paketleri yardımıyla oluşturulmuş; oluşturulan ortamlar aynı madde havuzu ve aynı bireyler üzerinden karşılaştırılmıştır. Çalışmada toplam 36 koşul, 30 replikasyon ile incelenmiştir. Simülasyonlar sonucunda, RMSE, yanlılık ve korelasyon değerleri hesaplanmış ve bu değerler aracılığıyla test desenlerinin performansları değerlendirilmiştir. Araştırma sonucunda, tüm test uzunlukları ve DMF oranı koşulları boyunca en iyi ölçüm hassasiyetini BBT deseninin sağladığı görülmüştür. ÇAT 1-3-3 ve 1-2-4 desenleri kıyaslandığında ise, 1-3-3 deseninin daha fazla koşulda yüksek ölçüm hassasiyeti sunduğu; ancak 1-2-4 desenine kıyasla daha yüksek performans gösterdiğini söylemek için bulguların yeterli olmadığı sonucuna varılmıştır. Ayrıca, DMF'li madde oranının artışından en az etkilenen desenin BBT deseni olduğu, ÇAT'ların ise özellikle 10 maddelik testte bu artıştan etkilendiği görülmüştür. Üç desen için de test uzunluğu arttıkça RMSE ve yanlılık değerinin düştüğü, korelasyon değerinin ise arttığı gözlenmiştir. Dolayısıyla madde sayısındaki artış ölçüm hassasiyetini de artırmıştır.

Anahtar sözcükler: bireyselleştirilmiş bilgisayarlı testler, çok aşamalı testler, değişen madde fonksiyonu

Abstract

The aim of this study was to investigate the effect of the presence of items indicating differential item functioning (DIF) on computer adaptive test approaches (CAT and MST) performance under several conditions. Both computer adaptive tests (CAT) and multi stage tests (MST) were investigated in terms of the fact that how they were affected by DIF items and how they perform under differents test length (10-20-30-40 item) and DIF item ratio (%10, %20, %30). For this purpose, 5000 persons' ability parameters and an item pool of 600 items were generated with simulation method and some of the items on the pool were made DIF items by changing b (item difficulty) parameters by the researcher. Then, one CAT and two MST environments (1-3-3 and 1-2-4 designs) were generated on RStudio by using 'xxIRT', 'catR' and 'mstR' packages. Those generated environments were compared over same individuals and item pool. In this study, 36 different condition were investigated with 30 replication. After simulation process; RMSE, bias and correlation values were calculated and those values were used in order to evaluate test designs's performances. As a result, it was seen that CAT has the best measurement precision over all test lengths and DIF ratio conditions. When MST 1-3-3 and 1-2-4 panel designs were compared, it was found that 1-3-3 design has higher measurement precision in more conditions; however, it was concluded that the findings were not enough to say that it performed better than 1-2-4 design. In addition to that, CAT was found to be the least affected design by the increase of DIF item ratio. MST designs were affected by that increament especially in 10-item length. As the test length increased it was seen that RMSE value decreased and correlation value increased for all test designs. Therefore, increase in test length resulted in an increase in the measurement precision.

Keywords: computer adaptive tests, multi stage tests, differential item funtioning

Teşekkür

Öncelikle tez çalışmamın başlangıcından bu yana yardımlarını esirgemeyen, fikirleriyle bana yol gösteren, bu süreçte her konuda destek sağlayan, güleryüzünü ve anlayışlı tavırlarını asla esirgemeyen, kendisiyle çalışma fırsatı bulduğum için oldukça şanslı hissettiğim değerli tez danışmanım Prof. Dr. Nuri DOĞAN'a;

Yüksek lisans ve doktora eğitimim boyunca çok şey öğrendiğim, bilgi ve deneyimlerinden faydalandığım, yalnızca akademik değil insani vasıfları ile de birçok konuda örnek aldığım değerli hocalarım Prof. Dr. Selahattin GELBAL, Prof. Dr. Hülya KELECİOĞLU ve Doç. Dr. Burcu ATAR'a;

Önerileri ve fikirleri ile tezime katkıda bulunan değerli jüri üyesi hocalarım Prof. Dr. Şeref TAN ve Doç. Dr. Hamide Deniz GÜLLEROĞLU'na;

Süreç içerisinde fikirleriyle çalışmama katkı sağlayan oda arkadaşlarım Osman TAT ve Abdullah Faruk KILIÇ'a ve süreçteki stresime katlanan, dertlerimi dinleyen tüm araştırma görevlisi arkadaşlarıma;

Her anımda yanımda olan, dertlerime ortak olan, hayatıma anlam katan ve tez sürecimdeki stresimi paylaşan Emine TUNÇ, Funda UYSAL, Gözde KAPLAN, Başak ÇİĞDEMTEKİN ve Zahide TONGA'ya;

Bugünlere gelmemde büyük emeği olan, her zaman arkamda ve destekçim olduklarını hissettiğim sevgili annem Rabia ERDEM'e, ablam Özlem ERDEM ÇAVDAR'a ve rahmetli babam Tuncer ERDEM'e;

Tüm süreçleri benimle birlikte yaşayan, her daim bana destek olan, yanımda olan ve her zaman olacağını bildiğim, anlayışı sonsuz sevgili eşim Hakan KARA'ya;

Doktora eğitimim süresince burs olanağı sağlayan TÜBİTAK'a;

Bu süreçte en ufak katkısı bulunan, adını sayamadığım herkese **ne kadar teşekkür etsem azdır.**

İçindekiler

Tablolar Dizini

Şekiller Dizini

Simgeler ve Kısaltmalar Dizini

BBT: Bireyselleştirilmiş Bilgisayarlı Test

ÇAT: Çok Aşamalı Test

DMF: Değişen Madde Fonksiyonu

OTB: Otomatik Test Birleştirme

RMSE: Root Mean Square Error - Hata Ortalamalarının Karekökü

TBF: Test Bilgi Fonksiyonu

TBDMF: Tek Biçimli Değişen Madde Fonksiyonu

TBODMF: Tek Biçimli Olmayan Değişen Madde Fonksiyonu

TEA: Tanımlanmış Evren Aralıkları

YMB: Yaklaşık Maksimum Bilgi

3 PLM: 3 Parametreli Lojistik Model

Bölüm 1
Giriş

Bu bölümde; problem durumu, araştırmanın amacı ve önemi, araştırma problemi, sınırlılıklar, sayıltılar ve tanımlar yer almaktadır.

Problem Durumu

Geleneksel lineer testler, 1900'lerden bu yana eğitimsel değerlendirmelerin temel taşı olmuştur. Genellikle kâğıt üzerinde uygulanan ve kâğıt-kalem testi olarak isimlendirilen bu testler bireylerin bilgi, beceri ve yeteneklerini ölçmede oldukça yaygın şekilde kullanılmıştır (Yan, Lewis ve von-Davier, 2017; Weiss ve Kingsbury, 1984). Ancak, özellikle son 40 yıldır bilgisayar teknolojisinde meydana gelen büyük değişimlerle birlikte, bilgisayar ortamında uygulanan testlerin popülerliği artmış ve bu testler lineer testlerin yerine uygulanabilir bir alternatif halini almıştır. Daha önce kâğıt-kalem formatında uygulanan birçok test, artık bilgisayar ortamında uygulanmaya başlanmıştır (Keng, 2008; Luecht ve Sireci, 2011; Magis ve diğerleri, 2017; Yan, von-Davier ve Lewis, 2014). Bilgisayar tabanlı testler (computer based tests); lineer testler, bireyselleştirilmiş bilgisayarlı testler (BBT) ve çok aşamalı testler (ÇAT) olarak üç başlıkta ele alınabilir (Magis ve diğerleri, 2017).

Bilgisayar tabanlı lineer testler, geleneksel lineer testlerin bilgisayar ortamında uygulanan halidir. Lineer testlerde olduğu gibi bu testlerde de tüm bireyler aynı maddeleri yanıtlar ve test uzunluğu sabittir. Dolayısıyla, ayrı bir madde havuzu gerektirmez ve test formları oluşturmak kolaydır. En büyük avantajlarından biri ise, içerik üzerinde test geliştiricilerin kontrolünün yüksek olmasıdır. Uygulama öncesinde geliştiriciler konu alanlarını, hangi alandan ne kadar soru sorulacağını belirleyebilir ve içerik kontrolünü sağlayabilir (Magis ve diğerleri, 2017; Sarı, 2016; Yan ve diğerleri, 2014). Öte yandan, bütün bireylere maddelerin kolaylık ve zorluk düzeyleri dikkate alınmaksızın aynı maddelerin sunulması bazı sorunlar oluşturabilmektedir. Lord (1980), doğru ölçüm için bir testin zorluk düzeyinin bireyin yetenek düzeyine uygun olması gerektiğini ifade etmiştir. Düşük yetenek düzeyindeki bireylerin zor maddeleri yanlış yanıtlamaları veya yüksek yetenek düzeyindeki bireylerin kolay maddelerin tümüne doğru yanıt vermeleri beklenen bir durumdur. Dolayısıyla, madde güçlüğü ile bireyin yetenek düzeyi arasındaki fark arttıkça maddelerin bireyin yeteneği hakkında sağladığı bilgi de azalmaktadır.

Sonuç olarak, geleneksel lineer testler grup içerisinde ortalama yetenek düzeyindeki bireylerin yeteneklerini en yüksek hassasiyette ölçebiliyorken, uçlardaki bireyler için daha az hassasiyette ölçümler sağlamaktadır. Bu testlerde hassas bir yetenek kestirimi elde edebilmek için çok sayıda madde gerekmektedir (Hambleton ve Swaminathan, 1991; Magis ve diğerleri, 2017; Wainer, 2000, 2000). Özetle, bilgisayar tabanlı lineer testler bilgisayarın getirdiği uygulama, puanlama, form oluşturma gibi kolaylıkların yanında geleneksel lineer testlerdeki aynı sınırlılıklara sahiptir (Magis ve diğerleri, 2017).

Bireyselleştirilmiş bilgisayarlı testlerde (BBT) ise temel amaç, testte gelen maddeleri her bireyin yetenek düzeyine göre yönlendirmek ve testin birey için çok kolay veya çok zor olmamasını sağlamaktır. Testi alan bireyin yeteneği süreç içerisinde kestirilerek, yetenek düzeyine uygun maddeleri yanıtlaması sağlanır (Tay, 2015; Yan ve diğerleri, 2014; Zheng ve Chang, 2014). Süreç içerisinde, birey bir maddeyi cevaplar ve bu noktada bireyin anlık yeteneği kestirilir. Kestirilen anlık yeteneğe bağlı olarak bir sonraki madde havuzdan algoritma ile seçilir ve bireye uygulanır. Sonlandırma kuralı karşılanana kadar, bireyin karşısına maddeler teker teker bu şekilde gelmeye devam eder (Tay, 2015; Weiss ve Kingsbury, 1984). BBT'lerin geleneksel lineer testlere kıyasla en temel avantajı, istenen ölçüm hassasiyetine daha az sayıda maddeyle ulaşılabiliyor olmasıdır (Wang, 2013; Wang, 2017). Bireyler yetenek düzeylerine uygun maddeleri aldıkları ve zamanlarını kendileri için daha zor veya daha kolay maddelerle uğraşarak kaybetmedikleri için, süreç daha kısa testlerle sonuçlanır (Wainer, 2000). Embretson ve Reise (2000), iyi desenlenmiş bir BBT uygulamasının, test uzunluğunu ölçüm doğruluğunu kaybetmeden %50'ye kadar azaltabildiğini belirtmiştir. Daha az sayıda madde kullanımı, bireylerin test sonuçlarını etkileyebilecek yorgunluk faktörünü de azaltmaktadır. Geleneksel testlerde ölçümlerin hassasiyeti bireylerin yetenek düzeylerine göre değişebilmekte iken; BBT'lerde bütün bireyler için eşit hassasiyette ölçümler sağlanabilmektedir. Yukarıda belirtildiği gibi, geleneksel testler için grup içerisinde ortalama yetenek düzeyine sahip bireylerin ölçümleri daha hassas yapılabilirken, uçlara yaklaştıkça ölçümlerin hassasiyeti azalmaktadır. BBT'lerde ise, uçlardaki bireyler de dâhil olmak üzere her birey için eşit hassasiyette ölçümler sağlanabilmektedir (Lord, 1980; Wang, 2017). Bilgisayarların kullanılıyor olması ise puan bildirimi, esnek uygulama planı, daha iyi test güvenliği kontrolü sağlama gibi

kolaylıklar sağlamaktadır (Lord, 1980; Yan ve diğerleri, 2014). Ancak, BBT'ler avantajlarının yanı sıra bazı dezavantajları da beraberinde getirmektedir. Bireylerin önceki maddelere dönerek gözden geçirme şansının olmaması bu testlerin en büyük dezavantajı olarak kabul edilmektedir. Ayrıca, içerik özelliklerini sağlamak ve madde kullanım sıklığı oranlarını kontrol etmek için karmaşık madde seçim yöntemlerine ihtiyaç duyulmaktadır (Hambleton ve Swaminathan, 1991; Hendrickson, 2007; Yan ve diğerleri, 2014).

BBT'lerin dezavantajlarının üstesinden gelebilmek ve aynı zamanda psikometrik etkililiğinden faydalanabilmek için, çok aşamalı testler (ÇAT) önerilmiştir (Hendrickson, 2007; Wang, 2017). ÇAT'larda, BBT'lerin ve lineer testlerin birçok avantajı bir araya getirilirken dezavantajları da minimuma indirilmektedir. Bu yönleriyle son yıllarda ÇAT'lar gittikçe popüler hale gelmiştir (Hendrickson, 2007; Magis ve diğerleri, 2017). Temel mantığı BBT'ye oldukça benzer olan bu testlerde, madde seçimleri tek tek değil de madde setleri (modül) şeklinde yapılmakta ve test aşamalar halinde yapılandırılmaktadır. Bir diğer deyişle, ÇAT'lar madde düzeyinde değil de madde setleri (modül) düzeyinde bireyselleştirilen testlerdir (Hendrickson, 2007; Yan, 2010). Birey bir modülü cevaplar, yeteneği kestirilir ve kestirilen yetenek düzeyine uygun başka bir modül seçilerek bireyin karşısına getirilir. ÇAT'larda, her bir modül içeriksel ve istatistiksel olarak istenen özelliklere sahip şekilde oluşturulabilir. Böylece, oluşturulmak istenen test formu üzerinde BBT'lere göre daha fazla kontrol sahibi olunabilir. ÇAT'lar, BBT'lere nazaran daha az uyarlama noktasına sahiptir; ancak daha etkili test oluşturma ve kontrollü içerik dengelemesi sağlamaktadır. Ayrıca, bireylerin aynı modül içerisinde önceki soruları tekrar gözden geçirebilme ve cevaplarını değiştirebilme şansı bulunmaktadır (Hendrickson, 2007; Wainer, 2000; Wang, 2017). BBT'lerde başlangıçta birey hakkında sahip olunan bilgi az olduğu için, testin başında verilen cevaplar kestirilen yetenekte büyük değişikliklere yol açmaktadır. Bu durum BBT'ler için yüksek yetenek düzeyindeki bireylerin erken hatalardan kurtulmalarını zorlaştırmaktadır. ÇAT'larda ise bireyin başlangıç yeteneği bir madde setinin cevaplanmasından sonra kestirildiği için bu durumdan daha az etkilenilmektedir (Rome, 2017). Sağladığı avantajların yanı sıra, ÇAT'ların olumsuz yönleri de bulunmaktadır. ÇAT'larda, BBT'lere nazaran daha fazla sayıda madde ile aynı ölçüm hassasiyetine ulaşılabilmektedir (Berger, Verschoor, Eggen ve Moser, 2019). Ayrıca, ÇAT'larda modüller yalnızca hedef

yetenek düzeylerinde (örneğin; düşük, orta ve yüksek yetenek düzeylerinde üçer düzey) en uygun güçlüklerde tasarlandığından, final yetenek kestirimleri BBT'ler kadar hassas olmamaktadır (Rome, 2017).

Bilgisayara dayalı test uygulamalarının artışı, özellikle testin adilliği noktasında bazı zorlukları da beraberinde getirmiştir (Chu ve Lai, 2013; Gierl, Lai ve Li, 2013; Zwick, 2010). Test adilliği ve eşitliği konuları, belirli öğrenci gruplarına karşı yanlılık gösteren maddelerle ilgilenmektedir. Yansız maddeler, testin ölçtüğü özellikle ilişkisiz faktörlerden (örneğin; cinsiyet, sosyo-ekonomik durum vb.) etkilenmeyerek yalnızca bireyin ölçülmek istenen özelliğini ölçmektedir. Yanlı maddeler ise, ölçülmek istenen özellikle ilişkisiz olan bu faktörlerden etkilenmektedir. Test sonuçlarının bireylerin geleceğini etkileyebilecek kritik karar alımlarında kullanıldığı da göz önünde bulundurulduğunda, testlerin adilliğinin önemi daha da ortaya çıkmaktadır (Camilli ve Shepard, 1994; Crocker ve Algina, 1986; Hambleton ve Swaminathan, 1991). Yanlılık konusunda bilgi elde etmek için kullanılan pek çok istatistiksel yöntem bulunmaktadır ve değişen madde fonksiyonu (DMF) analizleri en çok kullanılan yöntemlerden biridir. DMF analizleri ile yanlılık noktasında problemli olabilecek maddeler belirlenir ve sonrasında bu maddelerin gerçekten problemli olup olmadıkları noktasında uzman görüşüne başvurulur (Zumbo, 1999).

Bireyselleştirilmiş bilgisayarlı test yaklaşımlarında DMF. BBT uygulamalarının niteliği, büyük oranda madde havuzunun niteliğine/kalitesine bağlıdır (Han ve Guo, 2011). Bu uygulamalar için geniş madde havuzları yapılandırılmalı ve havuzdaki tüm maddelerin temel adillik ve eşitlik koşullarını sağladığı kontrol edilmelidir (Gierl, Lai ve Li, 2013). BBT'lerde kullanılan tüm maddelerin bu standartları karşıladığından emin olmak için, içerik uzmanlarının her bir maddeyi inceledikleri panelleri kapsayan duyarlılık incelemeleri (sensitivity reviews) yapılması önerilmektedir (Gierl, Lai ve Li, 2013; Zieky, 2006). Bu incelemelerde ayrıca DMF analiz sonuçları da gözden geçirilmekte, böylece uzmanlar yüksek DMF içeren maddelere odaklanabilmektedir (Gierl, Lai ve Li, 2013). Madde yazımının iyi yapılandırılması maddelerde DMF görülmesi ihtimalini azaltır. Ancak, DMF'den etkilenen birçok farklı alt grup olabildiği için DMF'nin etkilerinden tamamen kaçınmak mümkün olamamaktadır. Bireylerin bilgisayar aşinalığı, sınav ortamı, fiziksel yetersizlikler gibi maddeyle ilişkili olmayan pek çok

faktör DMF oluşturabilmektedir (Birdsall, 2011). Maddelerin içeriğinden bağımsız olarak, maddelerin sunuldukları bağlam da (örneğin; madde sıralaması) madde parametrelerini etkileyebilmektedir (National Research Council, 1999). Dolayısıyla bu durum da maddeler için bir DMF kaynağı olabilmektedir. Ayrıca, oluşturulan havuzdaki maddeler başlangıçta DMF göstermiyor olsa bile zaman içerisinde DMF gösterir hale gelebilmektedir. Zaman içerisinde maddelerin tekrar tekrar kullanımı sonucu, madde diğer bireyler tarafından uygulama öncesinde bilinir duruma gelebilmekte; bu durum yaşanmasa bile madde ile testi alan birey arasındaki etkileşimin değişimi zaman içerisinde çeşitli sebeplerle oluşabilmektedir. Bu durum madde parametre kayması (item parameter drift) olarak bilinmektedir. Böylelikle, madde ve birey arasındaki farklılaşan etkileşim temel olarak başlangıçta kalibre edilen madde özelliklerinden farklı made özelliklerine neden olur (Aksu-Dünya, 2017; Han ve Guo, 2011). Maddelerde gerçekleşen parametre kayması bir DMF türü olarak tanımlanmaktadır. Çünkü maddeler farklı test uygulamalarına katılan gruplar arasında farklı davranmaktadır (Aksu-Dünya, 2017; Babcock ve Albano, 2012). Madde parametre kayması test geçerliği ve adilliği için ciddi bir tehdit oluşturmaktadır (Han ve Guo, 2011).

Havuzda bulunan ve DMF gösterdiğinden şüphelenilen maddelerin gözden geçirilerek düzeltilmesi veya testten atılması önerilmektedir (Lei, Chen ve Yu, 2006). Ancak DMF'li maddelerin testten atılmasının iki dezavantajı bulunmaktadır. Birincisi çıkarılmak istenen maddenin aslında önemli bir unsuru farklı şekilde ölçebiliyor olması ve çıkarılmasının yapının ölçülmesinde boşluklar yaratabilecek olmasıdır. Bu durum da testin geçerliğinin düşmesiyle sonuçlanır. İkinci dezavantaj ise test çok sayıda koşul boyunca yapıldığında (örneğin; farklı diller) çok sayıda DMF'siz madde elde etmenin oldukça zor olmasıdır (Makransky ve Glas, 2013). Bu noktada, farklı alt gruplar için ayrı madde parametreleri kullanma yoluna gidilebilir. Örneğin; Bjorner, Chang, Thissen ve Reeve (2007) yaptıkları bir çalışmada madde havuzunda bir maddenin cinsiyet değişkeni için DMF gösterdiğini belirlemiş ve bu madde için kadın ve erkeklerde ayrı madde parametreleri kullanılarak bu durumun düzeltilebileceğini önermişlerdir.

DMF analizleri, bireyselleştirilmiş test yaklaşımları (BBT ve ÇAT) için lineer testlere nazaran daha önemli olabilmektedir. Bu testlerde daha az sayıda madde uygulandığı için, her bir madde bireylerin yetenek kestirimlerine daha fazla katkı

sağlamaktadır. Madde sayısının daha az olması dolayısıyla, bireylerin yanlı madde/ler alması durumunda, bu maddelerin ağırlığının yüksek olması ve yetenek kestiriminde daha önemli rol oynaması beklenebilir. Ayrıca, yanlılığın varlığı maddelerin uygulanma sıralarına da etki edebilir; çünkü BBT ve ÇAT'larda sıradaki madde/modül, önceki maddeye/modüle verilecek cevaplara göre belirlenmektedir (Zwick, 2010). ÇAT için yanlı maddelerin belirli modüllerde yoğunlaşması da sorun oluşturabilir.

Yukarıdaki bilgiler ışığında, yanlılık gösteren maddelerin testte yer almasının adil olmayan sonuçlar doğurabileceği ve hatalı kararlar alınmasına neden olabileceği varsayılmaktadır. Ayrıca, DMF'li maddelerin BBT'ler ve ÇAT'lar üzerindeki etkisinin lineer testlere nazaran daha fazla olabileceği belirtilmiştir (Zwick, 2010). Dolayısıyla, bu testlere ilişkin DMF çalışması yapılmasının önemli olduğu düşünülmektedir. İlgili literatür incelendiğinde, BBT ve ÇAT'larda yapılmış DMF çalışmalarının sınırlı (Chu ve Lai, 2013; Gierl ve diğerleri, 2013; Lei, Chen ve Yu, 2006; Piromsombat, 2014) olduğu görülmüştür. Ayrıca bu çalışmalarda, DMF belirleme yöntemlerinin etkililiğinin farklı koşullar altında sınanması (Chu ve Lai, 2013; Lei, Chen ve Yu, 2006; Gierl ve diğerleri, 2013) ve DMF gösteren maddelerin yetenek kestirimi üzerindeki etkisinin BBT'lerde incelenmesi ile (Piromsombat, 2014) sınırlı kalındığı belirlenmiştir. Testte DMF'li maddelerin bulunduğu durumda bireyselleştirilmiş bilgisayarlı test yaklaşımlarının karşılaştırmalı olarak incelendiği hiçbir çalışmaya rastlanmamıştır. Bu çalışmada, testte bulunan DMF'li maddelerin BBT ve ÇAT performanslarını nasıl etkilediği farklı koşullar altında incelenmiştir.

Araştırmanın Amacı ve Önemi

Bu çalışma kapsamında, testte DMF gösteren maddeler bulunduğu durumda BBT ve ÇAT performanslarının farklı koşullar (test uzunluğu, DMF'li madde oranı, ÇAT panel deseni) altında incelenmesi amaçlanmaktadır. Elde edilen sonuçlara dayanılarak minimum hataya ve yanlılığa, maksimum korelasyona sahip kestirimlerin hangi testten elde edildiği ve o test içerisinde hangi koşullarda sağlandığı konusunda fikir edinilebileceği düşünülmektedir.

Literatürde, BBT ve ÇAT'ların birçok yönünün (içerik dengeleme, madde havuzu özellikleri, test uzunluğu vb.) farklı koşullar altında karşılaştırıldığı çalışmalar bulunmaktadır. Ancak, her iki model için de DMF çalışmalarının oldukça sınırlı

olduğu görülmektedir (Zwick, 2010). Teste eklenen DMF'li maddelerin BBT ve ÇAT'lar üzerindeki etkisini karşılaştırmalı olarak inceleyen bir çalışmaya ise literatürde rastlanmamıştır. Bu çalışma kapsamında, DMF'li maddelerin BBT ve ÇAT üzerindeki etkisinin test uzunluğu, DMF'li madde oranı ve ÇAT panel deseni koşulları altında karşılaştırılmasından elde edilecek sonuçların literatüre katkı sağlayacağı düşünülmektedir. Ayrıca, Türkiye'de bugüne kadar ÇAT'a ilişkin yapılmış az sayıda (Boztunç-Öztürk, 2019; Doğruöz, 2018) çalışmaya rastlanmış olması ve BBT ve ÇAT'ların DMF yönünden incelenmemiş olması yönüyle çalışmanın ulusal ve uluslararası literatüre önemli katkılar sunacağı düşünülmektedir.

Araştırma Problemi

Bireyselleştirilmiş bilgisayarlı test (BBT) ve çok aşamalı test (ÇAT) uygulamalarında, testte DMF gösteren maddelerin bulunduğu durumda RMSE, yanlılık ve korelasyon değerleri;

- Testteki DMF'li madde oranına,

- Test uzunluğuna,

- ÇAT panel desenine göre

nasıl değişmektedir?

Alt problemler. Çalışma kapsamında aşağıdaki alt problemlere yanıt aranmıştır:

1. Testte bulunan DMF'li madde oranı %10 iken, farklı BBT yaklaşımlarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE, yanlılık ve korelasyon değerleri;

1a. Test uzunluğu 10,

1b. Test uzunluğu 20,

1c. Test uzunluğu 30,

1d. Test uzunluğu 40 madde olduğu durumda nasıl değişmektedir?

2. Testte bulunan DMF'li madde oranı %20 iken, farklı BBT yaklaşımlarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE, yanlılık ve korelasyon değerleri;

2a. Test uzunluğu 10,

2b. Test uzunluğu 20,

2c. Test uzunluğu 30,

2d. Test uzunluğu 40 madde olduğu durumda nasıl değişmektedir?

3. Testte bulunan DMF'li madde oranı %30 iken, farklı BBT yaklaşımlarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE, yanlılık ve korelasyon değerleri;

3a. Test uzunluğu 10,

3b. Test uzunluğu 20,

3c. Test uzunluğu 30,

3d. Test uzunluğu 40 madde olduğu durumda nasıl değişmektedir?

Sayıltılar

Veri üretimi aşamasında kullanılan madde parametrelerinin ve birey yetenek parametrelerinin dağılımlarının gerçek bir durumu yansıttığı varsayılmaktadır.

Sınırlılıklar

- Araştırma, simülasyon veri seti ile sınırlıdır.
- Araştırma, 1-0 puanlama verisi ile sınırlıdır.
- MTK modellerinden sadece 3 PLM kullanılmıştır.
- DMF'li maddeler yalnızca TBDMF gösterecek şekilde üretilmiştir.

Tanımlar

Bireyselleştirilmiş bilgisayarlı test yaklaşımı. Bireyin verdiği cevaplara göre karşılaşacağı soru seçiminin dinamik olarak yapıldığı bilgisayar tabanlı test modeli. Çalışma kapsamında hem BBT, hem de ÇAT yaklaşımlarını kapsayacak şekilde kullanılmıştır.

Bireyselleştirilmiş bilgisayarlı test (BBT). Bireyin her bir maddeye verdiği cevap sonrası yeteneğinin kestirildiği ve kestirilen yetenek düzeyine uygun maddeye yönlendirildiği test modeli.

Çok aşamalı test (ÇAT). Bireyin bir madde setine verdiği cevap sonrasında kestirilen yeteneği göz önünde bulundurularak bir sonraki cevaplayacağı madde setinin seçildiği test modeli.

Bölüm 2
Araştırmanın Kuramsal Temeli ve İlgili Araştırmalar

Bu bölümde, madde tepki kuramının bireyselleştirilmiş bilgisayarlı test yaklaşımlarında kullanımı hakkında genel bilgi sunulmuş, bireyselleştirilmiş bilgisayarlı testlerin ve çok aşamalı testlerin (ÇAT) genel kavramsal çerçevesi çizilmiştir. Ayrıca değişen madde fonksiyonu konusunda da bilgi verilmiş ve son olarak literatürde yer alan BBT'lerde değişen madde fonksiyonu ve BBT ve ÇAT karşılaştırması ile ilişkili örnek çalışmalara yer verilmiştir.

Madde Tepki Kuramı

Gözlenen değişkenler (örneğin; test maddeleri) ile örtük özellikler (örneğin; yetenek) arasındaki ilişkiyi tanımlamada yaygın olarak kullanılan iki ölçme kuramı; Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK)'dır. KTK, 20. Yüzyılın genelinde, test geliştirme için bir temel olmuş, yaygın bir şekilde kullanılmıştır. Ancak, Lord ve Novick (1968)'in çalışmalarıyla birlikte, Madde Tepki Kuramı psikolojik ölçme alanında önemli bir yer bulmuştur. KTK'nin uzantısı olarak türetilen MTK, uygulanan maddelerin özelliklerine ve bireylerin bu maddelere verdikleri cevaplara dayalı olarak yeteneklerinin kestirilmesi için geliştirilmiş olan modele dayalı bir yaklaşımdır (Embretson ve Reise, 2000). KTK özellikle teste ve gruba bağımlı sonuçlar üretiyor olması nedeniyle eleştirilirken; MTK parametre değişmezliği özelliğiyle bu sorunun üstesinden gelmiştir. MTK'ye göre, bireyin kestirilen yeteneği cevapladığı madde setine, maddelerin kestirilen özellikleri ise uygulandığı gruba bağımlı değildir. Bu özellikleriyle MTK popüler hale gelmiştir (Crocker ve Algina, 1986; Embretson ve Reise, 2000; Lord, 1980). Bu bölümde MTK'de kullanılan modeller ve terimlere ilişkin detaylı bilgi sunulmuştur.

İki kategorili MTK modelleri. Modele dayalı bir yaklaşım olan Madde Tepki Kuramı'nda birçok model kullanılmaktadır. Bu çalışma kapsamında, tek boyutlu-ikili puanlanmış veri üzerinde çalışıldığından iki kategorili MTK modelleri üzerinde durulmuştur. İki kategorili modeller, madde yanıtlarının iki kategorili şekilde (1: doğru, 0: yanlış) puanlandığı ikili verilerde kullanıma uygundur. Bu kapsamdaki en popüler modeller; bir, iki, üç ve dört parametreli lojistik modellerdir. Bu modeller, bireyin örtük yeteneği ile bir maddeyi doğru cevaplama olasılığı arasındaki ilişkiyi

modellemek amacıyla her bir madde için kestirilen madde parametresi sayısına göre isimlendirilir (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1991).

En basit iki kategorili MTK modeli olan bir parametreli lojistik model (1 PLM), maddeler arasında sadece güçlük parametresinin (b_i) değişimine izin verir ve ayırıcılık parametresi bütün maddelerde sabit kabul edilir. Yani, tüm maddelerin ayırıcılığı eşit olup, güçlük parametreleri farklılık göstermektedir. Şans parametresinin ise ihmal edilebilir düzeyde olduğu varsayılır (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1991). Rasch modeli ise, ayırıcılık değerinin tüm maddeler için 1'e eşit kabul edildiği, özel bir 1 PL modelidir. Bu modelde, bireyin bir maddeyi doğru cevaplama olasılığı ile yetenek düzeyi arasındaki ilişki güçlük parametresi üzerinden tanımlanır;

$$P\left(X_{ip}=1|\theta_p\right)=\frac{e^{(\theta_p-b_i)}}{1+e^{(\theta_p-b_i)}}$$

1 PLM'de olduğu gibi ayırıcılıkların eşit olduğu varsayımı, her bir maddenin farklı yetenek düzeyleri boyunca eşit düzeyde etkili olduğu anlamına gelmektedir. Ancak bu varsayım her zaman sağlanmamakta, eşit ayırıcılıkta olmayan maddeleri modellemek için 2 PLM'ye başvurulmaktadır (Lamoré, 2017). İki parametreli modele (2 PLM), 1 PLM'ye ek olarak ayırıcılık parametresi (a_i) eklenir. Bu modelde maddelerin güçlüklerinin yanı sıra ayırıcılık değerlerinin de farklılaştığı kabul edilmektedir. Ancak, 1 PLM'de olduğu gibi bu modelde de şans başarısı dikkate alınmamıştır (Hambleton ve Swaminathan, 1991). Bireyin bir maddeyi doğru cevaplama olasılığı aşağıdaki formül yardımıyla hesaplanır;

$$P\left(X_{ip}=1|\theta_p\right)=\frac{e^{a_i(\theta_p-b_i)}}{1+e^{a_i(\theta_p-b_i)}}$$

Üç parametreli modelde (3PLM) ise, güçlük ve ayırıcılık parametresine ek olarak şans başarısı parametresi de modele dâhil edilir. Birnbaum (1968) tarafından 2PLM'ye, şansın doğru cevaplama olasılığına katkısını simgeleyen bir şans parametresi (c_i) eklenmiş ve üç parametreli model elde edilmiştir. Baker (2001), bireylerin yalnızca tahminle de maddeleri doğru yanıtlayabileceklerini, doğru yanıtlama olasılığının küçük bir parçasının da şans başarısı olduğunu belirtmiştir. Şans parametresinin eklenmesiyle birlikte, bilgisi olmayan bir bireyin bile maddeyi

şans başarısıyla doğru cevaplayabileceği kabul edilir. Üç parametreli model için formül aşağıdaki gibidir;

$$P\left(X_{ip}=1|\theta_p\right)=c_i+(1-c_i)\frac{e^{a_i(\theta_p-b_i)}}{1+e^{a_i(\theta_p-b_i)}}$$

Son olarak, Barton ve Lord (1981) üç parametreli modele eklemeler yapmış ve dört parametreli modeli (4 PLM) ortaya koymuştur. Yüksek yetenekli bireyler, bazı durumlarda (örneğin; dikkatsizlik, kaygı vs) doğru yanıtlaması gereken maddelere yanlış yanıt verebilmektedir (Hambleton ve Swaminathan, 1991). Buradan yola çıkılarak, 4PLM'de oldukça yüksek yetenek düzeyindeki bireylerin maddeyi doğru cevaplama olasılıklarının 1'e eşit olmayabileceği olasılığı ele alınmıştır. Bunun için, diğer üç modelde 1 olarak alınan üst asimptot değerine ilişkin parametre (γ_i) 3PLM denklemine eklenmiştir. Dört parametreli model için kullanılan denklem aşağıda verilmiştir:

$$P\left(X_{ip}=1|\theta_p\right)=c_i+(\gamma_i-c_i)\frac{e^{a_i(\theta_p-b_i)}}{1+e^{a_i(\theta_p-b_i)}}$$

Aslında, her bir model bir diğer modelin alt kümesidir. Örneğin; 3 PLM, 4 PLM'nin üst asimptot değerinin 1'e eşitlendiği; 2PLM, 3 PLM'nin şans başarısının 0'a sabitlendiği özel bir halidir. 1PLM ise, 2PLM'nin ayırıcılık değerinin bütün maddeler için sabit olarak alındığı özel halidir.

MTK modellerinden faydalanılan başlıca alanlardan biri, bireyselleştirilmiş bilgisayarlı testlerdir (Ayala, 2009; Embretson ve Reise, 2000). Bu modeller, uygulanan belirli bir madde setinden bağımsız olarak yetenek kestirimleri sağlaması yönüyle BBT'ler için oldukça uygundur. Her birey farklı zorluk düzeyinde farklı maddeleri/madde setlerini cevaplamış olsa bile, bu modeller yardımıyla bireyler için yetenek kestirebilmek ve bireyler arası karşılaştırmalar yapabilmek mümkündür (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1991; Hambleton, Swaminathan ve Rogers, 1991).

Madde ve test bilgisi. MTK'de sıklıkla kullanılan terimlerden birisi de madde ve test bilgisidir. Bir yetenek düzeyine ilişkin ölçüm hassasiyeti (measurement precision) yetenek ölçeği boyunca aynı kalmamakta, madde parametre değerleri hassasiyet değerine etki etmektedir. MTK'de madde bilgisinden faydalanılarak bireylerin yetenek düzeylerinin ölçüm hassasiyetlerine karar verilebilmektedir. Belirli

bir yetenek düzeyindeki bilgi fonksiyonu ne kadar yüksek olursa, madde bireyin o düzeydeki yeteneğini de o kadar hassas şekilde ölçer (Embretson ve Reise, 2000; Yang, 2016). İkili puanlanan maddeler için madde bilgi fonksiyonu aşağıdaki eşitlik yardımıyla hesaplanır (Lord, 1980);

$$I_i(\theta_p) = a_i^2 \left[\frac{Q_i(\theta_p)}{P_i(\theta_p)}\right]\left[\frac{P_i(\theta_p) - c_i}{1 - c_i}\right]^2$$

Bu eşitlikte,

$I_i(\theta_p)$: i maddesinin θ_p yetenek düzeyinde sağladığı bilgi

a_i : i maddesinin ayırıcılığı

c_i: i maddesine ilişkin şans parametresi

$P_i(\theta_p)$: Bireyin i maddesini doğru yanıtlama olasılığı

$Q_i(\theta_p)$: Bireyin i maddesini yanlış yanıtlama olasılığı (1- $P_i(\theta_p)$)

Formülden de anlaşıldığı üzere, madde ayırıcılığının artması o maddenin sağladığı bilgi miktarının da artması anlamına gelmektedir.

Test bilgisi ise, tüm maddelerin sağladıkları bilgilerin toplamıdır. MTK'deki yerel bağımsızlık varsayımı gereği, test maddeleri ve dolayısıyla sağladıkları bilgi fonksiyonları birbirinden bağımsızdır. Sonuç olarak, tüm maddelerin sağladıkları bilgilerin toplamı test bilgi fonksiyonunu vermektedir (Embretson ve Reise, 2000; Thissen, 2000).

$$TI(\theta) = \sum_{i=1}^{K} I_i(\theta)$$

Formülden de anlaşılabileceği üzere, testte yer alan madde sayısı (K) arttıkça testin sağladığı bilgi miktarı da artmakta, uzun testler kısa testlere göre daha hassas ölçümler sağlamaktadır (Yang, 2016). Embretson ve Reise (2000) madde ve test bilgilerinin MTK'de oldukça önemli rol oynadığını, madde ve testin ne dereceye kadar performans gösterebileceği hakkında kritik bilgiler sağladığını belirtmiştir. Ayrıca test bilgi fonksiyonu, belirli bir yetenek düzeyindeki bilgi miktarı arttıkça, testin o düzeyde daha hassas ölçümler sağlayacağı bilgisini vermektedir. Bu bilgi aracılığıyla, test geliştiriciler belirli kesme noktalarında maksimum bilgi sağlayacak maddeleri seçebilir, çeşitli amaçlara yönelik testler tasarlayabilirler. Test bilgi fonksiyonunun tersi ise ölçmenin standart hatasını (ÖSH) vermekte ve ÖSH ile de

belirli yetenek düzeylerindeki ölçüm hassasiyeti değerlendirilebilmektedir (Yang, 2016).

Bireyselleştirilmiş Bilgisayarlı Testler

Yıllar boyunca, bireysel testler ve grup testlerinin kullanımı üzerinde süregelen tartışmalar olmuştur. Bireysel testlerde, birey yetenek düzeyine uygun sorularla karşılaşırken; grup testlerinde ise her birey için benzer bir durum sağlama ve sınav maliyetinin azaltılmış olması avantajları vardır. Geçtiğimiz yüzyıl boyunca tercih hep grup testlerinin kullanımı yönünde olmuştur (Wainer, 2000). Ancak grup sınavlarında test edilecek geniş bir yetenek ranjı bulunmaktadır ve her bireyin yeteneğini etkili bir şekilde ölçebilmek için testte bu ranja hitap edecek şekilde her düzeyde soru yer almalıdır. Eğer testte zor maddeler bulunmuyorsa, tüm maddeleri doğru yanıtlayan orta düzey ve üst düzey yetenekli bireyler arasında ayrım yapılamayabilir. Benzer şekilde, kolay maddelerin bulunmaması durumunda ise, orta düzeyli bireylerle düşük yetenek düzeyindeki bireyler arasında ayrım yapmak zorlaşabilir. Ayrıca, bireyler yetenek düzeylerine uymayan sorularla karşılaşmakta ve kendileri hakkında bilgi sağlamayan soruları yanıtlamaktadırlar (Lord, 1968; Wainer, 2000).

Lord'un (1971) çalışmalarıyla birlikte, grup sınavlarının daha esnek hale getirilebileceği düşüncesi ortaya çıkmıştır. Lord (1971), grup sınavlarının özellikle çok yüksek ve çok düşük yetenek düzeyine sahip birçok birey için etkili olmadığını görmüş ve bu sınavların bireyselleştirilmesi üzerine çalışmalar yürütmüştür. Bilgisayarlar yardımıyla birçok birey aynı anda aynı veya farklı testleri alabilmekte ve eğer istenirse, algoritmalar yardımıyla verilen madde havuzundan her bir birey için farklı testler oluşturulabilmektedir. Bir bireyin yeteneğini en etkili şekilde ölçmenin yolu, test maddelerinin bireyin yetenek düzeyine en uygun maddeler olması, birey için çok zor ya da çok kolay olmamasıdır. Bilgisayar algoritmaları yardımıyla, bireyin cevapladığı maddelere göre yeteneğinin kestirilmesi ve bu kestirimlere dayalı olarak bireye uygulanacak diğer maddenin belirlenmesi süreci izlenerek bireyselleştirilmiş testler oluşturulabilmektedir (Hambleton ve diğerleri, 1991; Lord, 1968).

Bireyselleştirilmiş testler; bireye uygulanacak test maddelerinin, bireyin önceki maddeye verdiği cevaba göre seçildiği testlerdir. Bu testler maddeler bireye

uygun zorluk düzeyinde seçilir. Sonuç olarak, test uygulama süreci boyunca her bireyin yeteneğine uyarlanmış testler elde edilir ve böylelikle testin birey için çok kolay veya çok zor olmasının önüne geçilir (Wainer, 2000; Weiss ve Kingsbury, 1984).

Bireyselleştirilmiş bilgisayarlı testlerin temel çalışma prensibi şu şekildedir (Lord, 1971);

- Başlangıçta bir madde (genellikle orta zorlukta) seçilir ve bireye uygulanır.
- Bireyin ilk maddeye verdiği cevaba göre, yetenek düzeyi bilgisayar algoritmaları tarafından kestirilir.
- Bir sonraki madde, kestirilen yetenek düzeyine en uygun olacak şekilde madde havuzundan seçilir.
- Birey soruyu doğru yanıtladığında daha zor bir soruyla, yanlış yanıtladığında daha kolay bir soruyla karşılaşır.
- Durdurma kuralı sağlanana kadar süreç bu şekilde devam eder.

Weiss ve Kingsbury (1984), BBT'lerin (a) Madde Tepki Modeli, (b) Madde Havuzu, (c) Başlama Kuralı, (d) Madde Seçim Yöntemi, (e) Yetenek Kestirim Yöntemi ve (f) Sonlandırma Kuralı olmak üzere altı ana bileşenden oluştuklarını belirtmiştir. Bu bileşenlere ilişkin bilgiler aşağıda sunulmuştur.

Madde tepki modeli. Madde Tepki Kuramı bağlamında bireyselleştirilmiş bilgisayarlı testlerde 1, 2, 3 veya 4 parametreli modeller kullanılabilmektedir. Kullanılacak modelin seçiminde maddelerin doğası (açık uçlu, çoktan seçmeli vb.) ve seçilen modelin veriye uygunluğu göz önüne alınmalıdır (Weiss ve Kingsbury, 1984). Bireyselleştirilmiş testler madde tepki kuramına dayalı olmak zorunda değildir; ancak, MTK kullanımı bu testlerin uygulanmasının etkililiği bağlamında kullanışlıdır (Weissman, 2014). BBT'lerde en yaygın kullanılan MTK modeli 3 PLM'dir (Green, Bock, Humphreys, Linn ve Reckease, 1984; Wainer ve Mislevy, 2000). Bu modelin tercih edilmesinin temel nedeni, 3 PLM'nin çoktan seçmeli madde verisiyle 1 PLM ve 2 PLM'ye göre daha iyi uyum sağlamasıdır. BBT'lerde şans başarısının etkisi, bireyin uygun olmayan zorluk düzeyindeki sorularla nadiren karşılaşıyor olması nedeniyle yüksek değildir. Ancak; BBT'lerde bireylerin maddeleri boş bırakmasına izin verilmemesi ve cevabı bilmediklerinde tahminde bulunmaya

zorlanmaları şans başarısının etkisini ortaya çıkarmaktadır. 3 PLM'de yer alan şans parametresi, bu modelin BBT'lerle uyumunu artırmıştır (Hambleton ve diğerleri, 1991; Wainer ve Mislevy, 2000).

Madde havuzu. Madde havuzu, bir BBT uygulamasında bireylere uygulanabilecek maddelerin tümüdür ve bireyselleştirilmiş bilgisayarlı testlerin bir gerekliliğidir. BBT uygulamalarında birey, havuzdan seçilen farklı madde/madde setlerinden oluşan bireyselleştirilmiş testler alır. Dolayısıyla, madde havuzunun kalitesi BBT'ler üzerinde oldukça etkilidir (Flaugher, 2000). Uygun madde havuzu büyüklüğüne ilişkin kesin bir sayı bulunmamakla birlikte 100 maddelik bir havuzun kabul edilebilir sonuçlar sağlayabildiği kabul edilmektedir (Weiss ve Kingsbury, 1984; Embretson ve Reise, 2000). Havuzdaki maddelerin zorluk düzeylerinin, her yetenek düzeyine hitap edecek çeşitlilikte olması bireyselleştirilmiş test uygulamaları için bir gerekliliktir. BBT'nin amacı maddeleri testi alan bireylerin yetenek düzeylerine uyarlamak olduğu için, madde havuzunda her yetenek düzeyinden (kolay, orta, zor) birçok madde bulunmalıdır. Ayrıca, en etkili ölçümler ayırıcılığı yüksek maddelerin kullanımıyla sağlanmaktadır (Flaugher, 2000; Weiss ve Kingsbury, 1984). Baker (2001), madde ayırıcılık parametresinin genellikle [0.5, 2] aralığında, güçlük parametresinin ise [-3, +3] aralığında değerler aldığını belirtmiştir. Keng (2008) ise, havuzdaki madde parametre dağılımlarının sınavın amacıyla ilişkili olması gerektiğini vurgulamıştır. Uygulanan sınavın amacı havuzdaki genel madde bilgi dağılımını yani havuz bilgi fonksiyonunu yetenek düzeyi boyunca etkiler. Ölçülen özelliğin bütün yetenek ölçeği boyunca eşit derecede ölçülmesi amaçlanan norm-dayanaklı testlerde havuz bilgi fonksiyonu için ideal olan, dikdörtgensel bir dağılım göstermesidir. Kriter dayanaklı testlerde ise amaç, bireyin özelliğini yetenek ölçeği boyunca bir veya daha fazla kesme noktasına göre ölçmektir ve havuz bilgi fonksiyonunun bu noktalarda tepe noktası oluşturacak şekilde dağılması idealdir (Keng, 2008).

Başlama kuralı. Başlama kuralı bileşeni, bireyin teste hangi düzey madde veya maddelerle başlayacağını belirleme süreciyle ilişkilidir. Yetenek kestirimi sürecindeki ilk adım, başlangıç yetenek düzeyinin kestirimidir. Testteki ilk maddeyi seçmek için bireyin başlangıç yetenek düzeyinin bilinmesine ihtiyaç duyulmaktadır. Belirleme aşamasında iki farklı yol bulunmaktadır. Birinci yolda, birey hakkında önsel bilgi kullanılır. BBT'lerde, farklı öğrenciler farklı düzeyde maddelerle teste

başlayabilir. Örneğin; bireyin yetenek düzeyinin yüksek olduğuna ilişkin bir bilgimiz varsa, uygulama zor sorularla (veya tersi durumda kolay sorularla) başlatılabilir. Birey hakkında herhangi bir bilgi sahibi olunmadığı durumda başvurulan ikinci yol ise, bir başlangıç θ değeri atanmasıdır. Bu değerin belirlenebilmesi için en yaygın kullanılan yöntem, testi alan popülasyona ilişkin ortalama yetenek düzeyini belirten parametre değerini başlangıç θ değeri olarak atamaktır. Yeteneğin standart normal dağılım gösterdiği kabul edildiğinde bu değer 0 olmaktadır (Keng, 2008; Thissen ve Mislevy, 2000; Weiss ve Kingsbury, 1984). Eğer test çok kısa değilse, başlangıç düzeyinin hatalı belirlenmesi sonuçlarda ciddi etkilere yol açmaz; ancak doğru belirlenmesi uygulanan soru sayısını azaltır. Bu nedenle bu düzeyin doğru belirlenmiş olması önem taşımaktadır (Thissen ve Mislevy, 2000).

Madde seçim yöntemi. BBT'nin geleneksel kâğıt kalem testlerinden farkı, bireyler için en uygun maddelerin seçilip uygulanmasıdır. Maddeler test sürecinde seçilmekte ve madde seçimi için bir algoritmaya ihtiyaç duyulmaktadır. Madde seçim yöntemi BBT'nin temel bileşenlerinden biridir. Kullanılan algoritma seçim kuralının performansını ve dolayısıyla elde edilen sonuçları etkilemektedir. Farklı madde seçim yöntemlerinin kullanımı, aynı birey için farklı maddelerin seçimiyle sonuçlanabilmekte ve farklı θ kestirimlerine yol açabilmektedir (Keng, 2008; Sarı, 2016; Wang, 2017). Madde seçiminde en yaygın kullanılan yöntemler Maksimum bilgi yöntemi ve Bayes yaklaşımıdır (Wang, 2017; Weiss ve Kingsbury, 1984). Bu yöntemler aşağıda detaylı şekilde açıklanmıştır.

Maksimum bilgi yöntemi. Maksimum bilgi yönteminde, anlık yetenek düzeyinde Maksimum Fisher Bilgisini sağlayan madde seçilir. İki kategorili (1-0) bir veride Maksimum Fisher Bilgisi aşağıdaki formül aracılığıyla hesaplanmaktadır.

$$I_i(\theta) = \frac{\left[\frac{\partial P_i(\theta)}{\partial \theta}\right]^2}{P_i(\theta)(1 - P_i(\theta))} = \frac{[P_i'(\theta)]^2}{P_j(\theta)(1 - P_i(\theta))}$$

$P_i(\theta)$: θ yetenek düzeyindeki bireyin i maddesine doğru cevap verme olasılığı

Tek boyutlu 3PLM'de iki kategorili için kullanılan formül şu hale gelir;

$$I_i(\theta) = \frac{D^2 a_i^2 (1 - c_i)}{(c_i + e^{Di(\theta - i)})(1 + e^{-Da_i(\theta - b_i)})^2}$$

D=1.7

a_i = i maddesinin ayırıcılık parametresi

b_i = i maddesinin güçlük parametresi

c_i = i maddesinin şans parametresi

Her seferinde en fazla bilgiyi sağlayan maddenin seçiliyor olması, sınavın etkililiğini artırmaktadır. Ayrıca, her bir uygulanan maddeden sonra yetenek kestirimi yapılmakta ve kestirilen yetenek düzeyinden sonra o yetenek düzeyi için (o yetenek düzeyindeki birey için) en fazla bilgiyi sağlayan madde seçilmektedir. Bu durumda, her birey farklı maddeleri almakta ve sınav süresince pek çok farklı yol (path) ortaya çıkmaktadır. Dolayısıyla test güvenliği oldukça yüksektir (Kingsbury ve Zara, 1989). Ancak, bu yöntem oldukça popüler olmasına karşın, yüksek ayırıcılık düzeyindeki maddelerin daha fazla bilgi sağlaması sebebiyle daha fazla seçiliyor olması dezavantajı vardır. En yüksek ayırıcılık gücüne sahip maddelerin genellikle en yüksek bilgiyi sağlaması nedeniyle sürekli seçilmesi test güvenirliğini ve geçerliğini etkileyebilmektedir (Hambleton, Jac ve Pieters, 2000; van der Linden ve Glas, 2010; Wang, 2017). Bu nedenle, bu yöntemin kullanımında dikkatli davranılmalıdır. Madde kullanım sıklığının kontrol edilmesi, bu durumun önüne geçmek için etkili bir yöntem olabilir. Bir diğer yöntem olarak, Hambleton ve diğerleri (2000) maddenin, ilgili yetenek düzeyinde maksimum bilgiyi sağlayan maddeler arasından rastgele seçilmesini önermişlerdir. Örneğin; ilgili yetenek düzeyinde en yüksek bilgiyi sağlayan dört madde belirlenip, bu maddeler arasından rastgele bir madde seçilebilir. Ayrıca, tüm madde havuzundan her seferinde en fazla bilgi veren maddenin tespit edilmesi süreci oldukça zaman alıcı istatistiksel hesaplamalar gerektirmektedir (Kingsbury ve Zara, 1989).

Bayes yaklaşımı. Bayes yaklaşımında (Owen, 1975), anlık yetenek kestirimine ilişkin beklenen sonsal dağılım varyansının minimize edilmesi amaçlanmaktadır (Weiss ve Kingsbury, 1984). Bu yönteme göre, her birey teste başlangıç yetenek düzeyine ilişkin önsel bilgi ile başlar. Cevaplanan her bir maddeden sonra, verilen cevap ve önsel dağılım verileri kullanılarak yeni bir yetenek kestirimi yapılır ve yetenek kestirimine ait sonsal dağılım oluşturulur. Bu yöntemde, Bayes sonsal varyansını en fazla azaltan (yetenek kestirimi hatasını en fazla azaltan) madde seçilmektedir. Madde havuzundaki her bir madde için sonsal varyans, bireyin anlık yetenek kestirimi ve madde parametreleri göz önüne alınarak hesaplanır. Sonrasında, sonsal dağılımın varyansını azaltan maddeler seçilir ve

madde seçimi varyans istenen düzeye ulaşana kadar devam eder (Weiss ve Kingsbury, 1984). Maksimum Fisher Bilgisi'nde olduğu gibi, madde havuzunun büyüklüğü arttıkça en uygun maddeyi bulmak için harcanan zaman da artar ve test süreci yavaşlar (Kingsbury ve Zara, 1989). Chang ve Stout (1993), Bayes ve maksimum bilgi yaklaşımlarının başlangıç aşamalarında farklı madde seçimleriyle sonuçlanabildiğini ancak test uzunluğu arttıkça benzer sonuçlar verdiğini belirtmiştir. Ayrıca, madde seçim yöntemlerinin performanslarının karşılaştırıldığı çalışmalarda, maksimum bilgi yöntemi ile hiçbir yöntemin farklılaşmadığı görülmüştür (Veldkamp, 2003). Bu nedenle bu çalışmada maksimum bilgi yönteminin kullanılmasına karar verilmiştir.

Yetenek kestirim yöntemi. BBT'lerin avantajlarından biri, bireyin yetenek düzeyiyle uyumlu maddelerin seçilmesidir. Uygulanan her bir maddeden sonra bireyin yeteneği kestirilir ve bir sonraki madde kestirilen yetenek düzeyine uygun olacak şekilde yönlendirilir. Bu nedenle yetenek kestirim yöntemleri BBT'lerin önemli bileşenlerindendir. En yaygın kullanılan iki yöntem maksimum olabilirlik kestirimi ve Bayes kestirim yöntemleridir (Embretson ve Reise, 2000; Keng, 2008; Sarı, 2016; Wang, 2017). Bu yöntemler detaylı bir şekilde aşağıda açıklanmıştır.

Maksimum olabilirlik kestirimi. Maksimum olabilirlik kestirimi (MOK) yöntemi, bir bireyin cevap örüntüsünün olabilirliğini maksimum yapan değeri bulmaya dayalıdır. Bir diğer deyişle, bireyin bir madde setine verdiği 0 ve 1'lerden oluşan cevap örüntüsü ve madde parametreleri bilindiği durumda, bireyin örtük yetenek sürekliliğindeki (latent-trait continuum) en muhtemel konumu araştırılmaktadır (Embretson ve Reise, 2000).

Olabilirlik fonksiyonu *L(θ),* bireyin yetenek düzeyi θ iken, verilen bir cevap örüntüsünün gözlenme olasılığıdır. Olabilirlik fonksiyonu ile verilen cevap örüntüsünün gerçekleşme ihtimali kestirilir ve örüntünün ortaya çıkma ihtimalini en yüksek yapan θ değeri kestirilmeye çalışılır. Yani, bireyin madde cevap örüntüsünün olabilirliğini maksimize eden nokta bulunur ve bu değeri maksimum yapan θ parametresi kestirilmeye çalışılır (Lord, 1980; Wang ve Vispoel, 1998). MOK yönteminde kullanılan olabilirlik fonksiyonu aşağıdaki gibidir (Wang, 2017);

$$\mathrm{L}(u|\theta) = \prod_{i=1}^{n} P_i(u_i|\,\theta, a_i, b_i, c_i)$$

n: madde sayısı

P_i (u_i | θ, a_i, b_i, c_i): i maddesinde u_i cevabını alma olasılığı

Elde edilen olabilirlik fonksiyonunun birinci türevinin 0'a eşit olduğu nokta bulunarak bireyin yeteneği kestirilir. Buradan kestirilen yetenek değeri (θ) maksimum olabilirlik değeridir.

$$\frac{\partial}{\partial\theta} L(u|\theta) = 0$$

Bu yöntemin etkili ve tutarlı sonuçlar verdiğinin belirtilmesiyle birlikte, en büyük dezavantajı bütün maddeleri doğru veya yanlış yanıtlayan bireylere ilişkin kestirim yapılamamasıdır. 0 puan alan bireyler için kestireceği yetenek -∞, tam puan alan bireyler için kestireceği yetenek +∞ olacaktır. Bu durum, özellikle bireyselleştirilmiş bilgisayarlı testlerin ilk aşamaları için tehlike arz eden bir durumdur ve test uzunluğunun kısa olduğu durumlarda kullanımı önerilmez. Bir doğru veya bir yanlış yanıt alana kadar MOK'un kullanımı tavsiye edilmemektedir (Hambleton ve Swaminathan, 1991; Wang, 2017).

Bayes kestirim yöntemi. Bayes yönteminde yetenek kestirimi için, olabilirlik fonksiyonu ile birlikte önsel bilgiden yararlanılır. Test uygulanmadan önce bireyin yetenek dağılımına ilişkin bilinenleri temsil eden önsel bilginin olabilirlik fonksiyonuyla harmanlanarak kullanımını sağlaması nedeniyle bu yöntemin daha etkili kestirimler sağladığı düşünülmektedir (Embretson ve Reise, 2000). Bayes yöntemleri, maksimum olabilirlik kestiriminde ortaya çıkan sonsuzluk sorununu çözmek için de bir alternatif olarak düşünülmektedir; çünkü bireylerin yeteneği ilk cevaplarından sonra kestirilebilmektedir (Wang, 2017). Bu yöntemde, testi alan bireylerin karşılaştıkları ilk maddeye verdikleri cevaba ilişkin olabilirlik fonksiyonu oluşturulur ve elde edilen fonksiyon önsel dağılımla harmanlanarak sonsal dağılım elde edilir. Elde edilen bu sonsal dağılım, bir sonraki maddeye verilecek cevabın önsel dağılımı olarak kullanılır (Wang ve Vispoel, 1998). Önsel dağılım Bayes teoremine göre güncellenerek bireyin yeteneğine ilişkin sonsal dağılım elde edilmektedir.

$$f(\theta|u) = \frac{f(u|\theta)\, f(\theta)}{f(u)}$$

f (θ|u) : Sonsal dağılım
f (θ) : Önsel dağılım
f ($u|\theta$): Verilen cevabın olabilirliği

Bayes yöntemlerinde bireyin yeteneği sonsal dağılımın merkezi eğilim ölçülerinden yararlanılarak kestirilir. EAP (Beklenen Sonsal-Expected a Posteriori) kestirimi sonsal dağılımın ortalamasını kullanırken, MAP (MAP-Maximum a Posteriori-Maksimum Sonsal) kestiriminde dağılımın modundan faydalanılmaktadır. Bayes yöntemleri MOK'un neden olduğu soruna çözüm bulmuş olsa da önsel dağılım seçiminin final yetenek kestirimi üzerinde etkisi olabilmektedir (Wang, 2017). Wang ve Vispoel (1998), uygun olmayan bir önsel dağılımın seçilmesi durumunda final kestiriminin yanlı olabileceğini belirtmiştir. Bu nedenle, önsel dağılımın belirlenmesi noktası oldukça önemlidir.

Bayes yöntemlerinin her ikisi de kısa testler için oldukça tutarlı ve daima sınırlıdır. Böylece, MOK'ta var olan bireylerin tüm maddeleri doğru veya yanlış yanıtlaması durumunda yetenek kestirimi yapamama sorunu bu yöntemlerde yoktur. Ancak bu yöntemler de bazı durumlarda yüksek yetenekleri olduğundan daha düşük veya düşük yetenekleri daha yüksek kestirme eğilimi göstermektedirler. Madde sayısı fazla değilse bu durum bir sorun olarak ortaya çıkabilmektedir (Keng, 2008). Bireyselleştirilmiş testlerde madde seçimi ve yetenek kestirimi için genellikle kullanılan kombinasyon madde seçiminde maksimum bilgi yöntemi ile birlikte yetenek kestiriminde EAP kestirimidir (van der Linden, 2008; van der Linden ve Pashley, 2010).

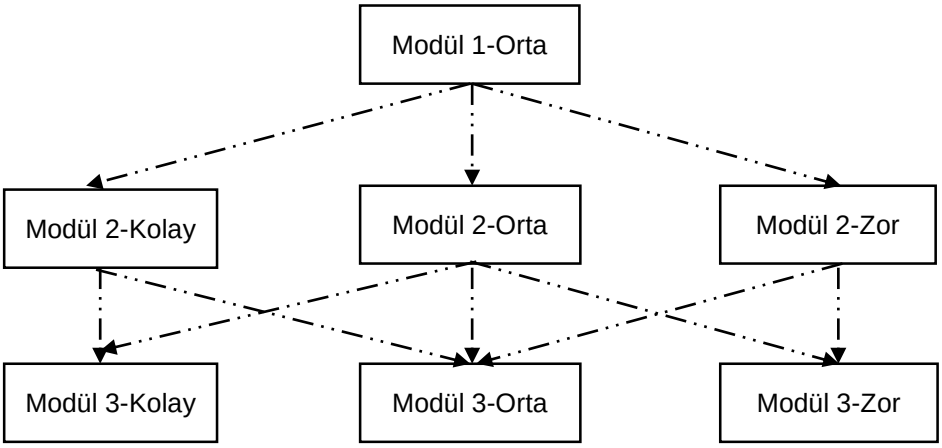
Sonlandırma kuralı. BBT uygulamalarında, testin ne zaman sonlanacağına karar verilebilmesi için bir sonlandırma kuralına ihtiyaç duyulmaktadır. Sabit uzunluk ve değişen uzunluk, testin ne zaman sonlandırılacağına karar vermekte kullanılan iki yoldur. Sabit uzunluklu BBT'lerde bütün bireyler aynı sayıda madde alırlar. Örneğin, belirlenen madde sayısı 30 ise, 30 maddeden sonra ölçüm doğruluğu/hassasiyeti düzeyine bakılmaksızın test sonlandırılır. Değişen uzunluktaki BBT'lerde ise yetenek düzeyi için önceden belirlenmiş hassasiyet (precision) düzeyine ulaşılana kadar test devam etmekte, dolayısıyla bireylerin cevapladıkları madde sayıları farklılaşabilmektedir. Belirlenmiş bir 'ölçümün standart hatası' değeri, sonlandırma kriteri olarak kullanılabilir. Böylece, her bir birey için aynı hassasiyet düzeyinde ölçüm yapılabilir (Keng, 2008; Sarı, 2016; Thissen ve Mislevy, 2000; Wang, 2017).

Çok Aşamalı Testler

Özellikle son yıllarda oldukça popüler hale gelmeye başlayan çok aşamalı testler (ÇAT), bireyselleştirilmiş bilgisayarlı testlere (BBT) alternatif olarak önerilmiştir. Yeni bir fikir olmayan çok aşamalı testlere ilişkin Angoff ve Huddleston (1958) ve Lord (1971) tarafından yapılmış, kağıt-kalem testlerinin çok aşamalı test şeklinde desenlendiği çalışmalar bulunmaktadır. Lineer testlerle BBT'lerin avantajlarını bir araya getiren ÇAT'ların BBT'lerden farkı, bireye uyarlamanın madde düzeyinde değil de madde setleri (modül) düzeyinde yapılmasıdır (Hendrickson, 2007; Yan ve diğerleri, 2014).

ÇAT, önceden oluşturulmuş madde gruplarının algoritmalarla seçildiği ve testin aşamalar şeklinde oluşturulduğu algoritmaya dayalı bir test yaklaşımıdır. ÇAT'ların kendine özgü modül, aşama, panel, rota (pathway) gibi bazı unsurları bulunmaktadır (Yan ve diğerleri, 2014). Test uygulamasından önce oluşturulan madde grupları/setleri *modül* olarak adlandırılır. ÇAT'larda öncelikle her bireye testin birinci *aşamasında* bir başlangıç madde seti uygulanır ve bu set *yönlendirme modülü* olarak adlandırılır (Yan ve diğerleri, 2014). Her bir aşamada farklı güçlük düzeylerinde modüller bulunmaktadır ve yönlendirme modülündeki cevaplara bağlı olarak kestirilen yetenek düzeyine göre birey bir sonraki aşamada kendisine en uygun yetenek düzeyindeki modüle yönlendirilir. Son yetenek kestirimi, testin tümüne verilen cevaplara dayalı olarak yapılır (Wang, Haiyan, Chang ve Douglas, 2016).

Genel anlamda ÇAT'ın çalışma prensibi şu şekildedir; Uygulama, yönlendirme modülüyle başlar. Bireyin yeteneği, yönlendirme modülündeki cevaplarına bağlı olarak kestirilir ve bu yetenek kestirimine göre birey ikinci aşamadaki uygun modüle yönlendirilir. Örneğin yönlendirme modülünde bireyin başarısı yüksekse, ikinci aşamada zor modülle karşılaşır. İkinci aşamanın tamamlanmasından sonra bireyin yeteneği tekrar kestirilir ve 3. aşamadaki uygun modüllerden birine yönlendirilir. Bu süreç, birey tüm aşamaları tamamlayıncaya kadar devam eder (Sarı, 2016). Şekil 1'de 1-3-3 desenli bir ÇAT örneği sunulmuştur.

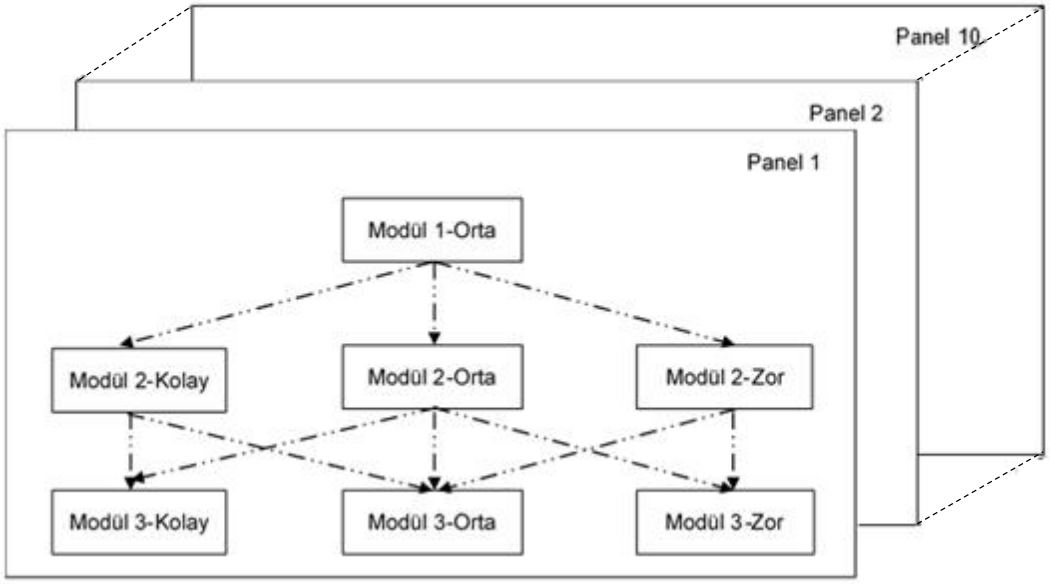


Şekil 1. 1-3-3 desenli çok aşamalı test örnek şeması

1-3-3 deseni, ilk aşamada bir, ikinci ve üçüncü aşamalarda ise üçer modül bulunduğu anlamına gelmektedir. Kolay, Orta, Zor ifadeleri ise modüllerin zorluk düzeylerini belirtir. Bu panel desenine sahip bir testi alan birey, öncelikle 'Modül 1-Orta' modülündeki maddeleri cevaplandırır. Bu modüldeki performansına bağlı olarak birey, 2. Aşamadaki modüllerden birine yönlendirilir. Birey iyi bir performans sergilemesi durumunda 'Modül 2-Zor' modülüne, orta düzeyde bir performans sergilediğinde 'Modül 2-Orta' modülüne, düşük düzeyde bir performans gösterdiğinde ise, 'Modül 2-Kolay' modülüne yönlendirilir. Aynı süreç 3. Aşama için de devam eder. Bireyin test süreci boyunca izlediği modül sırası ise 'rota' (pathway) olarak adlandırılır. Her birey sadece bir tane rota izler (Yang, 2016). Örneğin, 'Modül 1-Orta', 'Modül 2-Zor' ve 'Modül 3-Orta' modüllerini alan bir öğrencinin rotası 'Modül 1-Orta – Modül 2-Zor – Modül 3-Orta'dır. Luecht ve Nungester (1998) düşük düzeyden yüksek düzeye gerçekleşecek aşırı performans değişikliklerinin çok olası olmadığını, bu nedenle bireylerin kolay modülden zor modüle veya zor modülden kolay modüle yönlendirilmemesi gerektiğini belirtmiştir. Örneğin; 'Modül 2-Kolay' modülüne yönlendirilen birey bu modelde çok iyi performans sergilese bile 'Modül 3-Zor' modülüne yönlendirilmemelidir.

Modüller oluşturulduktan sonra gruplandırılarak, panel olarak isimlendirilen test uygulama birimleri elde edilir. Her bir rota için oluşturulan ve geleneksel kâğıt-kalem testlerindeki test formlarına karşılık gelen bu panel sayılarının artırımı sayesinde havuzdaki madde kullanım sıklığı kontrol altında tutularak test güvenliği artırılır. Çoklu panel kullanımı; panel, modül ve madde kullanım sıklığını ve maddelerin çok fazla sayıda kullanımını azaltmaya yardımcı olması nedeniyle

önemlidir. Sınavın önemine ve amacına bağlı olmakla birlikte, panel sayısı genellikle 1 ile 40 arasında değişmektedir (Sarı, 2016; Rome, 2017; Yan ve diğerleri, 2014). Üç aşamalı 10 panelli bir örnek ÇAT uygulaması Şekil 2'de sunulmuştur.



Şekil 2. 1-3-3 desenli 10 panelli bir ÇAT örneği

Şekil 2'de görüldüğü üzere modüller bir araya getirilerek her biri üç aşama içeren ve hepsi birbirine paralel olan on farklı panel oluşturulmuştur.

Diğer test modellerinde olduğu gibi, ÇAT'ın de avantaj ve dezavantajları bulunmaktadır. ÇAT, bireyselleştirilmiş testlerin özel bir hali olarak düşünülebilir ve lineer testlere kıyasla yetenek ölçeği boyunca daha etkili ve hassas ölçümler sağlamaktadır. Ayrıca, ÇAT'da uygulanan modüllerin uygulama öncesinde oluşturulması ve dizayn edilmesi yönüyle test geliştiriciye içerik dengelemesi, testin yapısı ve uygulanması noktasında daha fazla kontrol şansı tanınır. Bireylere her bir modül içerisinde önceki maddeleri gözden geçirme imkanı sunması, ÇAT'ların bir diğer önemli özelliğidir (van Der Linden ve Glas, 2010; Yan ve diğerleri, 2014). ÇAT'ta test uzunluğu, BBT'lerde olduğundan daha uzun ancak lineer testlere göre daha kısadır (Hendrickson, 2007). BBT'lerdeki yüksek maliyetli madde havuzu oluşturma sorunu, ÇAT için de önemli bir sorundur. Ölçüm etkililiği noktasında ise, ÇAT'lar lineer testlerden çok daha doğru ölçümler sağlarken, BBT ile eşit veya biraz daha düşük doğrulukta ölçümler sağlamaktadır (Hendrickson, 2007; Patsula, 1999; Sarı, 2016).

Bir ÇAT uygulaması oluşturulurken, madde havuzu, panel yapısı, test birleştirme (test assembly), yönlendirme stratejisi ve puanlama yöntemi gibi unsurlar göz önünde bulundurulmalı ve planlanmalıdır. Ayrıca; testin amacı, içerik özelliği ve bireylerin yetenek dağılımları bu planlamaları yaparken göz önünde bulundurulmalıdır (Wang, 2017). Her bir ÇAT'ta maddelerin veya madde setlerinin nasıl seçildiğini kontrol eden bir yönlendirme algoritması, testin oluşturulduğu bir madde havuzu ve anlık ve final yetenek kestirimini hesaplayan bir yetenek kestirim metodu bulunmalıdır. Yetenek kestirim yöntemi, MTK modeli gibi ÇAT'larda da ihtiyaç duyulan unsurlar yukarıda BBT bölümünde anlatıldığındanbu kısımda tekrar bu unsurlara yer verilmemiştir. ÇAT'a özgü diğer unsurlar aşağıda kısaca açıklanmıştır.

Test birleştirme. Lineer testlerde maddeler bir araya getirilerek test bilgi fonksiyonları veya diğer kriterler benzer olduğunda paralel olduğu düşünülen formlar oluşturulurken, ÇAT'larda ise rotalar lineer test formlarına eşdeğerdir. Ancak bir panelin kendi içerisindeki farklı rotalar farklı güçlük düzeylerini gösterdiği için asla paralel olmamaktadır. İki ÇAT paneli, karşılıklı gelen tüm rotaların paralel olması durumunda paralel olarak kabul edilir. Rotaların paralel olması bile, bireysel modüllerin mutlaka paralel olmasını gerektirmemektedir (Zheng ve diğerleri, 2014).

ÇAT'larda, test güvenliğinin artırılması ve havuzdaki maddelerin etkili kullanımı gibi nedenlerle çoklu paneller oluşturulmaktadır. Madde seçimi ile çoklu modül ve panellerin oluşturulması için madde havuzu ve matematiksel algoritma ile aynı anda istatistiksel ve istatistiksel olmayan test özelliklerine odaklanan süreç test birleştirme sürecidir. Bu süreçte, otomatik test birleştirme (automated test assembly) var olan bir madde havuzundan modülleri oluşturmak için etkili bir yöntemdir (Wang, 2017; Zenisky ve Hambleton, 2014; Zheng, Wang, Culbertson ve Chang, 2014). Luecht (2003) de otomatik test birleştirmeyi (OTB) içerik ve diğer nitel özelliklerle alakalı birçok sınırlamayı da göz önüne alarak, bir veya birden fazla test formu için madde havuzundan madde seçiminde matematiksel optimizasyon süreçlerinin kullanılması olarak tanımlamıştır. OTB'nin optimum çözümü; modüllerdeki ve panellerdeki maddelerin güçlük düzeyi, içerik kontrolü, kelime sayısı, madde formatı vb. istenen sınırlılıkların sağlandığı çözümdür (Sarı, 2016).

Maddelerin modüllere atanmasında yaygın olarak kullanılan iki yöntem; lineer programlama ve sezgisel (heuristic) yöntemlerdir (Wang, 2017). Lineer

programlama yöntemleri, tüm test birleştirme kısıtlılıklarını (içerik kapsamı, madde kullanım sıklığı vb.) kesin olarak karşılayan çoklu paralel paneller oluşturmak için en uygun çözümü sunar. Ancak sınırlılıkların sayısı fazla olduğunda süreç kullanışsız, zaman alıcı ve hatta imkânsız bir hal alabilmektedir. Öte yandan sezgisel yöntemler istenen tüm sınırlılıkların karşılanacağı garantisini vermemekte ancak daha az hesaplama süreci içermekte ve daima bir çözüm sunabilmektedir (Zheng, Nozawa, Zao ve Chang, 2012). Lineer programlama modeli çoklu paralel test formlarının oluşturulmasında esnekliği ve içerik alanı, madde türü, kelime sayısı gibi farklı test özelliklerine sahip madde havuzlarına uygulanabilirliği nedeniyle yaygın şekilde kullanılmaktadır (Zheng ve diğerleri, 2014). Karmaşık tamsayı programlama (Mixed integer programming – MIP), lineer programlama yöntemlerinin bilinen bir formudur (Wang, 2017).

Luecht ve Nungester (1998) tarafından modüllerin panellere atanması için önerilen iki yaklaşım ise ‘Yukarıdan-aşağıya’ (Top-Down) ve ‘Aşağıdan-yukarıya’ (Bottom-Up) yaklaşımlarıdır. Aşağıdan-yukarıya yaklaşımında, paralellik her bir modül için paralel formların oluşturulmasıyla sağlanır. Her bir modül, modül düzeyindeki içerik gerekliliklerini ve istatistiksel sınırlılıkları sağlayacak şekilde birbirinden bağımsız olarak oluşturulur ve bu şekilde oluşturulan modüller tamamen paralel oldukları için aynı güçlük düzeyindeki modüller paneller arasında değiştirilebilirdir. Her bir modülün alternatif formları paralel olduğu için ilişkili rotalar da otomatik olarak paralel olacaktır (Sarı, 2016; Wang, 2017; Xiong, 2018; Zheng ve diğerleri, 2014). Yukarıdan-aşağıya yaklaşımında ise, modül düzeyinde değil tüm test düzeyinde test özellikleri gereklidir ve modüller belirli şekillerde bir araya getirilerek test düzeyindeki amaçlara ulaşılır (Xiong, 2018). Bu yaklaşımda modüller test düzeyindeki kriterlere göre oluşturulmakta ve tamamen paralel olmayan modüller birleştirilerek paneller elde edilmektedir. İlk olarak, modüller paralellik göz önünde bulundurularak veya bulundurulmayarak oluşturulur ve sonrasında panel düzeyinde paralelliği ve istatistiksel olmayan sınırlılıkları sağlamak için ek bir optimizasyon turu yapılır. Dolayısıyla bu yöntemde modüller paralel olmadığı için paneller arasında değiştirilebilir (birbirinin yerine kullanılabilir) değildir (Wang, 2017; Zheng ve diğerleri, 2014). Yukarıdan-aşağıya yöntemi ÇAT için test birleştirme işlemini karmaşık hale getirmektedir. Bu yaklaşımda öncelikle panel oluşturulmakta ve sonrasında panel önceden tanımlanmış kuralları karşılayan belirli sayıda modüle

bölünmektedir. Genel olarak, aşağıdan-yukarıya yaklaşımı madde havuzu ve sınırlılıklar uygun olduğunda uygulanması daha kolay olan yaklaşımdır (Xiong, 2018; Zheng ve diğerleri, 2014)

Madde havuzu. Test içeriği ve hem istatistiksel hem de istatistiksel olmayan sınırlılıkları birleştiren madde havuzu, ÇAT'ta etkili ölçümler alabilmek için oldukça önemli görülmektedir. Hendrickson (2007) madde havuzunun, modülleri ve çoklu panelleri oluşturabilmek için yeterli büyüklükte olması gerektiğini belirtmiştir. Özellikle ÇAT için desenlenmiş bir madde havuzu kullanmak, puanlama doğruluğuna katkı sağlar (Wang, 2017). İdeal madde havuzu, istenen hedef zorluk ranjına sahip olabilmek, modül ve paneller için esneklik sağlayabilmek için yeterli büyüklükte olmalıdır. Wang (2017), ÇAT madde havuzları için uygun madde sayısının test için gerekli madde sayısının 1.5 katı olmasını önermiştir.

Panel yapısı. ÇAT uygulamasına başlamadan önce, panel desenine ve yapısına karar verilmesi gerekmektedir. Bu kapsamda aşama sayısı, her aşamadaki modül sayısı ve her modüldeki madde sayısı belirlenmelidir.

Aşama sayısı. ÇAT'larda olası aşama sayısı, iki ile toplam madde sayısı aralığında değerler alabilmektedir (Hendrickson, 2007). Kâğıt-kalem testlerinin uyarlamalı şekilde yapıldığı önceki çalışmalarda sadece iki aşamalı ÇAT kullanılmışken (Lord, 1971), bilgisayar tabanlı testlerdeki gelişmelerle birlikte son zamanlardaki uygulamalarda 3 veya 4 aşama da kullanılmaktadır (Hendrickson, 2007). Aşama sayısı, istenen içerik kapsamı ve ölçüm hassasiyeti göz önünde bulundurularak karar verilmesi gereken bir faktördür (Zenisky, Hambleton ve Luecht, 2010). Aşama sayısının artması sadece daha fazla uyarlama şansı ve daha düşük ölçüm hatası şansı tanıyabilir. Örneğin; ortalamanın üstünde bir öğrenci, beklenmeyen şekilde ilk aşamada kötü performans sergilerse, ikinci aşamada kolay modüle yönlendirilir. Eğer test iki aşamalı ise, elde edilen sonuç bireyin gerçek yeteneğini yansıtmayacaktır. Üçüncü aşama eklendiği takdirde bireye, daha bilgilendirici bir modüle yönlendirilmesi konusunda bir şans daha tanınmış olur (Wang, 2017). Ancak, teste fazla aşama eklemenin final test formlarının ölçüm hassasiyetine çok fazla katkıda bulunmaksızın, test formu oluşturmayı daha karmaşık hale getireceği de göz önünde bulundurulmalıdır (Hendrickson, 2007; Luecht ve Nungester, 1998). Birçok ÇAT araştırması ve uygulamasında; iki, üç ve dört aşamalı desenler kullanılmıştır. İki aşamalı testlerde sadece bir adaptasyon

(uyarlama) noktası olmasından kaynaklı olarak yönlendirme hatasının gerçekleşme ihtimali –özellikle de yönlendirme modülünden kesme noktasına yakın puan alan bireyler için- daha yüksektir. Bu nedenle, daha fazla aşama kullanılması bu hatanın azaltılmasını sağlayabilir (Yan ve diğerleri, 2014). Patsula ve Hambleton (1999) araştırmalarında aşama sayısının ikiden üçe çıkarılmasının yetenek kestirimindeki hata miktarını azalttığını ortaya koymuştur (aktaran Yan ve diğerleri, 2014). Zenisky ve diğerleri (2010) de, birçok araştırmanın iki ve üç aşamalı teste odaklandığını ancak iki aşamalı testlerin kullanıldığı durumlarda bazı bireylerin yanlış şekilde yönlendirildikleri durumun onarılamayabileceği ihtimaline karşı dikkatli olunması gerektiğini belirtmişlerdir.

Modül sayısı. Aşama sayısıyla benzer şekilde, bir aşamadaki modül sayısının artırılması da ölçüm hassasiyetini etkilemekte, daha fazla uyarlama şansı sunmaktadır. Her bir aşamada farklı zorluk düzeylerinde daha fazla sayıda modül bulunması, testi daha geniş bir yetenek ranjı için daha uyarlanabilir hale getirir (Hendrickson, 2007; Wang, 2017). Ancak, daha uyarlanabilir aşama modülleri oluşturmak daha fazla sayıda kolay ve zor madde anlamına gelir ve madde havuzu oluşturulması noktasında sıkıntı yaratabilir (Zenisky ve diğerleri, 2010). Çoğu çalışma ve uygulamada, ilk aşamada yönlendirme modülü olarak tek bir modül kullanımı tercih edilmekte ve sonraki aşamalarda kullanılan modül sayısı artırılmaktadır (Hendrickson, 2007). Patsula ve Hambleton (1999), farklı panel desenlerini karşılaştırdıkları çalışmalarında, aşama 2 veya aşama 3'te modül sayısını üçten beşe çıkardıkları durumda, yetenek kestirimlerinin doğruluğunun ve uçlardaki yetenek düzeylerinde ÇAT'ın, BBT ve lineer testlere göre etkililiğinin arttığını belirtmiştir. Ancak bu şekilde desenlerin karmaşıklığı da artmıştır (Aktaran Yan ve diğerleri, 2014). Önceki araştırmalar, üç aşamadan fazlasının ve bir aşama içerisinde dört modülden fazlasının puanlama doğruluğuna çok az bir farklılık kattığını ve test oluşturma karmaşıklığını artırdığını belirtmiştir. Genel anlamda, maksimum 4 modül ve 3 aşamanın yeterli olabileceği düşünülmektedir (Armstrong, Jones, Koppel ve Pashley, 2004; Zenisky ve diğerleri, 2010).

Madde sayısı. Araştırmalarda ve uygulamalarda bir modülde kullanılan madde sayısı 1'den 90'a kadar farklılaşmakta, ortalama madde sayısı ise 5 olmaktadır (Armstrong ve diğerleri, 2004; Hendrickson, 2007; Yan ve diğerleri, 2014). Modüllerdeki madde sayıları aşamalar arasında farklılık gösterebilir. Bazı

testlerde yönlendirme modülünde madde sayısı fazlayken sonraki aşamalarda test uzunluğu daha kısa olabilmektedir. Kim ve Plake (1993), yönlendirme modülündeki madde sayısını artırmanın, yetenek kestirim hatalarını azaltmada önemli bir etkisi olduğunu belirtmiştir. Patsula ve Hambleton (1999) ise yetenek düzeylerinin çoğunluğunda modüllerdeki madde sayılarının aşamalar arasında farklılık göstermesinin yetenek kestirimlerinin doğruluğu üzerindeki etkisinin çok az olduğunu belirtmiştir. Ayrıca bu farklılığın, lineer testlerle ve BBT'lerle karşılaştırıldığında, ÇAT'ların göreli etkililiği üzerinde oluşturduğu etkinin de çok az olduğu ifade edilmiştir (Aktaran Yan ve diğerleri, 2014).

Yönlendirme yöntemi. ÇAT'deki yönlendirme yöntemi, BBT'deki madde seçim yöntemine benzerdir. ÇAT'ın amacına ve desenine dayalı olarak oldukça farklılık gösterebilen yönlendirme yöntemi, bireyleri önceki aşamadaki performansına dayalı olarak farklı rotalara veya bir sonraki aşamadaki modüllere, seçili kuralları kullanarak yönlendirir. Yönlendirme yönteminin etkililiği, öğrencinin izleyeceği rotayı ve dolayısıyla da yanlış yönlendirme durumunda testin kullanışlılığını etkiler. Bu nedenle ÇAT'ın önemli bir parçasıdır (Sarı, 2016; Yan ve diğerleri, 2014). Hendrickson (2007), kullanılabilecek iki yöntemin doğru cevap sayısı ve MTK'ye dayalı MOK ve EAP gibi yetenek kestirimleri olduğunu belirtmiştir.

Yönlendirme kesme noktalarının belirlenmesi ise 'Yaklaşık Maksimum Bilgi' (YMB) (Approximate Maximum Information-AMI) veya 'Tanımlanmış Evren Aralıkları'(TEA) (Defined Population Intervals - DPI) yöntemleriyle yapılabilir. YMB yönteminde modül seçimi için optimal karar noktasını belirlemede toplam test bilgi fonksiyonundan faydalanılırken, TEA panel ve modül yapısı boyunca orantısal rotalar çizmek için kullanılır. YMB yönteminde, daha önce uygulanan modüllere dayalı toplam test bilgi fonksiyonu ve alternatif modüllere (örneğin; kolay, orta, zor) ilişkin test bilgi fonksiyonları hesaplanır. Sonrasında, her bir alternatif modüle ilişkin TBF, toplam TBF'ye ayrı ayrı eklenir ve elde edilen TBF'lerin kesişim noktaları kesme noktası olarak tanımlanır (Luecht, Brumfield ve Breithaupt, 2006). Örneğin, 1-3-2 deseninde birinci aşamadan ikinci aşamaya geçiş için kesme noktaları belirlenmek istendiğinde, öncelikle birinci aşamada bireyin cevaplandırdığı modüle ait TBF hesaplanır. Daha sonra ikinci aşamadaki üç ayrı modül (2K-2O-2Z) için ayrı ayrı TBF'ler hesaplanır. Elde edilecek iki kesme noktasından ilki, 1 + 2K ve 1 + 2M'nin, diğeri ise 1+2M ve 1+2Z'nin kesişim noktasıdır. TEA yönteminde ise, birey

popülasyonunun belirli bir oranının bir sonraki aşamadaki farklı modülleri alması beklenir ve bireyler bu şekilde yönlendirilir (Luecht ve diğerleri, 2006). Örneğin; eğer 1-3-3 bir MST'de, ikinci aşamadaki modüllerin her birine bireylerin 1/3'ünün gitmesini istiyorsak, yetenek dağılımında 33. ve 67. yüzdeliklere karşılık gelen değerler belirlenerek kesme noktası olarak atanır.

BBT ve ÇAT'lar yukarıda anlatılan ana unsurların yanı sıra, test güvenliğini artırmak için kullanılan madde kullanım sıklığı da önemli bir unsurdur.

Madde kullanım sıklığı. BBT'lerde madde havuzlarında hem kaliteli hem de zayıf maddeler yer almaktadır. Bu testlerde amaç ilgili yetenek düzeyinde birey için en kaliteli (ilgili yetenek düzeyinde en fazla bilgiyi sağlayan) maddeyi seçmek olduğundan, iyi maddelerin seçilme olasılığı yüksekken kötü maddelerinki düşüktür ve iyi maddeler farklı bireyler için tekrar tekrar seçilirken kötü maddeler kullanılmadan kalır (Sarı, 2016). İyi maddelerin sık sık seçiliyor olması bireylerin bazı maddeleri hatırlamasına ve sonraki bireylere iletmesine imkân sağlayabilmekte ve bu durum test güvenliği için önemli bir tehdit oluşturmaktadır. Eğer madde haddinden fazla kullanılırsa, bu maddeyle alakalı ön bilgi ulaşılabilir hale gelir ve bazı bireyler arasında adaletsizlik yaratır. Aynı zamanda, madde havuzu geliştirmek zaman alıcı ve maliyetli bir süreçtir ve kaliteli her bir maddeden faydalanılmak istenmektedir (Rudner, 2010; Sarı, 2016; Wang, 2017). Bu nedenlerle madde kullanım sıklığının kontrol altında tutulması önemli bir husustur. BBT'lerde madde kullanım sıklığı kontrolü için kullanılan en yaygın ve popüler yöntemlerden birisi 'seçkisizlik' (randomesque) (Kingsbury ve Zara, 1989) yöntemidir. Bu yöntemde, ilgili yetenek düzeyinde maksimum bilgiyi sağlayan maddeyi seçmek yerine optimum düzeydeki madde grubu arasından bir madde rastgele seçilir. Örneğin; teste her seferinde θ =0 yakınlarında en fazla bilgiyi sağlayan maddeyle başlamak yerine, en iyi üç madde arasından rastgele seçilen maddeyle başlanabilir. Bu yöntemin uygulaması kolay olmakla birlikte, kullanım oranının istenen düzeyle sınırlı olması garanti edilmez (Keng, 2008; Magis ve diğerleri, 2017; Sarı, 2016; Wang, 2017). ÇAT'larda ise madde kullanım sıklığının kontrolü çoklu panellerin kullanımıyla mümkün olmaktadır. Panel olarak adlandırılan paralel test formlarının oluşturulması ve bireylerin bu panellere rastgele atanması ile panel, modül ve madde kullanım oranlarının azaltılması sağlanır (Sarı, 2016; Yan ve diğerleri, 2014)

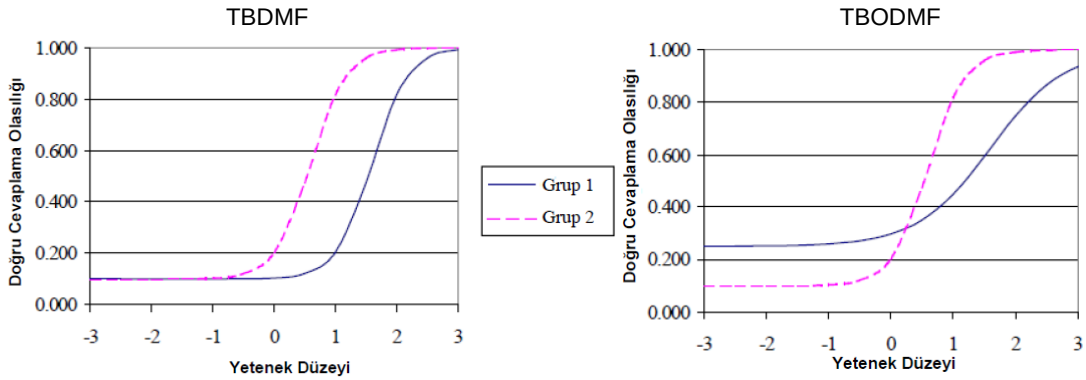
Değişen Madde Fonksiyonu

Günümüzde test sonuçları birçok karara temel oluşturması yönüyle büyük önem taşımaktadır. Ancak, testlerden elde edilen sonuçlar ölçülmek istenen özellik haricinde farklı varyans kaynakları tarafından etkilenebilmektedir. Test puanlarının bu kaynaklardan etkilenmesi kaçınılmaz olduğundan, en azından bir alt gruba yönelik avantaj sağlanmamasına dikkat edilmelidir. Değişkenlik kaynaklarının alt popülasyonları farklı şekilde etkilemesi yanlılık göstergesidir. Yanlılık, testteki maddelerin, ölçülmek istenen özellikle ilişkisiz bir özellikten kaynaklı olarak, alt gruplardan birinin lehine veya aleyhine çalışmasıdır (Camilli ve Shepard, 1994; Crocker ve Algina, 1986).

Maddelerin yanlılık gösterip göstermediğinin belirlenebilmesi için farklı yöntemler bulunmaktadır. Bunlardan biri, maddelerin değişen madde fonksiyonu (DMF) gösterip göstermediğinin belirlenmesidir. Hambleton ve diğerleri (1991) DMF'yi aynı yetenek düzeyindeki iki farklı grubun bir maddeyi doğru cevaplama olasılıklarının farklılaşması olarak tanımlamıştır. Camilli ve Shepard (1994), DMF'nin ortaya çıkmasına yol açabilecek iki olası nedenin alt gruplar arasındaki gerçek yetenek farklılığı (madde etkisi) ve madde yanlılığı olduğunu belirtmiştir. Dolayısıyla, DMF'nin var oluşu yanlılığın değil, gruplar arası gerçek yetenek farklılığının göstergesi olabilir. Bu nedenle, DMF analizleri eşit yetenek düzeyindeki bireyler arasında gerçekleştirilmelidir (Clauser ve Mazor, 1998). DMF'nin varlığı, ölçülmek istenen davranış haricinde bir özelliğin madde performansını etkilediğine, maddenin bireysel fark varyasyonlarının birden fazla boyutunu yansıttığına işaret etmektedir. Böyle bir testten çıkarılan sonuçların geçerliği tehlikeye girmektedir; çünkü elde edilen sonuçlar testin ölçmek istediği özellik haricinde farklı niteliklerin göstergesi olabilmektedir (Steinberg, Thissen ve Wainer, 2000). DMF gösterdiği belirlenen madde uzmanlar tarafından incelenir ve uzmanlardan madde fonksiyonu farklılaşmasının kaynağının ölçülen yapıyla ilişkili olup olmadığına karar vermesi beklenir. Yani, DMF tek başına bir yanlılık göstergesi değildir.

MTK perspektifinden bakıldığında, DMF grupların madde tepki fonksiyonları arasındaki fark olarak tanımlanabilir. İki grubun madde karakteristik eğrileri özdeşse maddenin DMF göstermediği sonucuna varılabilir. Eğrilerin birbirinden önemli derecede farklı olduğu durumda ise, madde DMF gösteriyordur (Zumbo, 1999).

DMF, tek biçimli ve çok biçimli olmak üzere iki farklı şekilde ortaya çıkabilmektedir. Bir maddenin tüm yetenek düzeyleri boyunca aynı grup lehine çalıştığı durumda *tek biçimli DMF* (TBDMF); madde, belirli yetenek düzeylerinde bir grup lehine işlerken farklı yetenek düzeylerinde diğer grup lehine işlediğinde *tek biçimli olmayan DMF* (TBODMF) ortaya çıkmaktadır (Hambleton ve diğerleri, 1991). Şekil 3'te her iki durumu da gösteren madde karakteristik eğrileri sunulmuştur.



Şekil 3. TBDMF ve TBODMF gösteren madde karakteristik eğrileri

Şekil 3'te görüldüğü gibi, TBDMF'de madde karakteristik eğrileri (MKE) kesişmemekte, madde tüm yetenek düzeyleri boyunca grup 2 lehine çalışmaktadır. MKE'lerin çakıştığı TBODMF grafiğinde ise, alt yetenek düzeylerinde grup 1 lehine çalışan madde, üst yetenek düzeylerinde grup 2 lehine çalışmaya başlamıştır. TBDMF'de MKE'ler hiçbir zaman çakışmazken, eğrilerin kesişmiş olması da TBODMF'nin bir göstergesidir (Zumbo, 1999).

DMF analizleri uyarlamalı testler için oldukça önemli bir yere sahiptir. Uyarlamalı testlerde bireylerin yanıtlayacakları maddelerin seçimi bir önceki maddede gösterdikleri performansa bağlıdır. Dolayısıyla DMF içeren bir maddenin uygulanması kendisinden sonra gelecek maddeyi değiştirebilir veya maddelerin uygulanma sırasını farklılaştırabilir. Ayrıca, uyarlamalı testlerde test uzunluğu lineer testlere kıyasla daha az olmakta, bu durum DMF içeren maddelerin yetenek kestirimine etkisini artırabilmektedir (Zwick, 2010; Gierl ve diğerleri, 2013; Zwick ve Bridgeman, 2014). Son olarak, test uygulamasının bilgisayar aracılığıyla yapılması bilgisayar aşinalığı, kaygı, ortam gibi geleneksel testlerde bulunmayan bazı olası DMF kaynaklarını ortaya çıkarabilmektedir (Zwick, 2010). Steinberg ve diğerleri (2000), bireyselleştirilmiş testlerin, DMF'nin geçerlik üzerindeki etkilerine karşı lineer

testlere kıyasla daha hassas olabildiğini belirtmiştir. Bu faktörler, BBT'lerde DMF analizlerinin önemini artırmıştır.

İlgili Araştırmalar

Bu bölümde, BBT ve ÇAT bağlamında DMF'nin incelendiği ve BBT ve ÇAT uygulamalarının karşılaştırıldığı örnek çalışmalara yer verilmiştir.

BBT yaklaşımları bağlamında DMF çalışmaları. Bu bağlamda yapılmış DMF çalışmaları daha çok BBT uygulamalarında DMF'li maddelerin doğru tespit edilmesi ve tespit etmekte kullanılan yöntemlerin etkililiği üzerine odaklanmıştır. Lei, Chen ve Yu (2006), CATSIB, CAT-LR ve CAT-IRTLR yöntemlerinin DMF belirleme performanslarını karşılaştırdıkları bir simülasyon çalışması yapmıştır. Farklı örneklem büyüklüğü oranlarının ve test etkisi büyüklüğünün koşul olarak alındığı bu çalışmada, kullanılan üç yöntemin tek biçimli DMF'yi belirleme gücü bakımından benzer sonuçlar verdiği tespit edilmiştir. Ancak, tek biçimli olmayan DMF'yi belirleme noktasında; CAT-LR ve CAT-IRTLR, CATSIB yöntemine göre daha güçlüdür.

Gierl, Lai ve Li (2013), bir ÇAT bağlamında madde havuzuna DMF gösteren maddeler eklemiş ve DMF tespiti için CATSIB yönteminin performansını farklı koşullar altında değerlendirmiştir. DMF tespit oranlarını etkileyebileceği düşünülen üç koşul; madde güçlüğü, örneklem büyüklüğü ve dengelenmiş/ dengelenmemiş desen manipüle edilmiştir. Elde edilen sonuçlar, CATSIB'ın sadece orta büyüklükte (475+; 175 odak grup, 300 referans grup) ve büyük örneklemlerde (600+; 300er birey odak ve referans grup) tutarlı ve doğru sonuç verdiğini göstermiştir.

Piromsombat (2014) tarafından yapılan çalışmada ise, operasyonel maddelerde DMF bulunmasının yetenek kestirimi üzerindeki etkisi BBT'ler üzerinde incelenmiştir. Ayrıca, bireylerin BBT'den elde edilen yetenek kestirimlerine, bilgisayar tabanlı lineer bir sınavdan elde edilen yetenek kestirimlerine ve doğru cevap sayılarına göre eşleştirilmesinin DMF'li maddeleri belirleme doğruluğu üzerindeki etkisi karşılaştırılmıştır. Bu çalışmadan elde edilen sonuçlar, DMF'li maddelerin testin başlarında gelmesi durumunda, özellikle de DMF düzeyi orta düzeyde olduğunda, BBT'nin DMF etkisini düzenleyebildiğini göstermiştir. Diğer durumlarda, BBT DMF'nin etkisini azaltmış ancak yetenek kestirimini DMF etkilerinden yeterince koruyamamıştır. Bu çalışma, testte bulunan DMF'li

maddelerin BBT'ler üzerindeki etkisinin incelenmesine yönelik literatürde rastlanan tek çalışmadır.

BBT ve ÇAT desenlerinin karşılaştırıldığı çalışmalar. BBT ve ÇAT karşılaştırmalarında; içerik dengeleme, madde havuzu özellikleri, panel desenleri vb. yönlerden uygulamaların etkililiğinin incelendiği çalışmalar bulunmaktadır. Bu başlık altında desen karşılaştırmalarına ilişkin örnek çalışmalara yer verilmiştir. Kim ve Plake (1993), BBT'nin ve ÇAT'ın ölçüm hassasiyetini ve göreli etkililiğini ilk aşama modül uzunluğu (10, 15 ve 20 madde), toplam test uzunluğu (40, 45 ve 50 madde), ikinci aşama modül sayısı (6, 7, 8 modül) ve ilk aşama modülündeki madde güçlüğü dağılımı koşullarında araştırmıştır. Elde edilen bulgular, BBT'nin ÇAT'a göre hem ölçüm hassasiyeti hem de göreli etkililik noktalarında daha iyi sonuç verdiğini ortaya koymuştur.

Patsula (1999) tarafından yapılan çalışmada; farklı BBT, kağıt-kalem testi ve ÇAT desenlerinden (aşama sayısı, her aşamadaki modül sayısı, her modüldeki madde sayısı) elde edilen yetenek kestirimlerinin doğrulukları kıyaslanmıştır. Çalışma sonuçlarına göre, BBT'ler en doğru yetenek kestirimini üretmiş; aşama sayısı ile modül sayısının artırılması ise ÇAT sonuçlarını BBT sonuçlarına yaklaştırmıştır. ÇAT'larda aşama sayısının ikiden üçe çıkarılması yetenek kestirimindeki hata miktarını azaltmıştır. Benzer şekilde, modül sayısının üçten beşe çıkarılması yetenek kestirimlerinin doğruluğunu artırmıştır.

Kim, Chung, Dodd ve Park (2012), BBT ve ÇAT'a yönelik farklı test desenlerini karma testler (mixed-format) üzerinde sınıflama testi bağlamında incelemiştir. ÇAT'de birinci aşama modülü, test bilgi fonksiyonu için üç farklı merkez alınarak üç düzeye göre oluşturulmuş, ayrıca üç farklı geçme oranı koşul olarak eklenmiştir. İlk aşama modülündeki TBF düzeyleri yüksek olduğunda, daha yüksek sayıda doğru sınıflama oranı elde edilmiştir. BBT için maksimum bilgi koşulunun kullanıldığı durumda, en yüksek sınıflama doğruluğu elde edilmiş, seçkisizlik yönteminin (randomesque-10) kullanıldığı BBT ise, artan TBF düzeyleriyle birlikte ÇAT'la karşılaştırılabilir sonuçlar vermiştir. Ayrıca, tüm ÇAT koşulları için ÇAT en iyi havuz kullanım oranlarını vermiş, BBT'nin iki desenine göre de daha iyi test güvenliği sağlamıştır. Ortalama maksimum madde kullanım oranları ÇAT, BBT maksimum bilgi ve BBT randomesque-10 için sırasıyla .337, .843 ve .322 olarak hesaplanmıştır.

Zheng, Nozawa, Zao ve Chang (2012), Kim ve diğerleri (2012)'ne benzer şekilde bir sınıflama testi geliştirmiş, otomatik test birleştirme yöntemi kullanarak ÇAT tasarlamış ve oluşturulan ÇAT'ın performansını lineer formuyla ve BBT haliyle kıyaslamışlardır. BBT ve ÇAT kıyaslandığında, ölçüm hassasiyeti ve madde havuzu kullanımı bakımından elde edilen sonuçların benzer olduğu gözlenmiştir. Ancak sınıflama açısından bakıldığında; ÇAT deseni BBT kadar iyi derecede sınıflama doğruluğu sunmuş ve bu esnada daha etkili bir madde havuzu kullanımı sağlamıştır.

Bir diğer çalışmada ise Sarı (2016), farklı uzunluktaki testlerde içerik alanları sayısı değişmekte iken BBT ve ÇAT'tan elde edilen sonuçların hassasiyetini araştırmıştır. Test uzunluğu (24 madde ve 48 madde), içerik alanı sayısı ve panel deseni (1-3 ve 1-3-3) çalışma kapsamında manipüle edilen özelliklerdir. Çalışma sonucunda test uzunluğu ve test deseninin sonuçları içerik alanı sayısından daha fazla etkilediği görülmüştür. Çalışmanın ana bulgusu, herhangi bir çalışma koşuluna bakılmaksızın BBT'nin diğer iki ÇAT desenine göre daha iyi sonuç verdiği, iki ÇAT deseninin ise karşılaştırılabilir sonuçlar sunmuş olduğudur.

Wang (2017) ise, BBT ve ÇAT için ayrı desenlenmiş farklı madde havuzları altında ölçüm doğruluğu ve ortalama test uzunluğunu karşılaştırmıştır. ÇAT için 16 farklı desen koşulu (1-2-3 ve 1-3-3 panel deseni; YMB ve TEA yönlendirme stratejileri; 45 ve 60 maddelik test uzunlukları; yukarıdan-aşağıya ve aşağıdan-yukarıya test birleştirme yöntemleri) manipüle edilmiştir. Her bir koşul ilişkili olduğu BBT sonuçlarıyla karşılaştırılmıştır. Edinilen bulgulardan yola çıkılarak, ÇAT ve BBT'nin benzer ölçüm hassasiyeti sağladığı sonucuna varılmıştır. BBT uzunluğu orta uzunlukta (moderate-length) olduğunda, yukarıdan aşağıya birleştirilmiş üç aşamalı bir ÇAT desenine geçiş makul ve uygulanabilirdir; ancak uzun testler için BBT kullanımı önerilmiştir.

Yurtiçinde yapılan çalışmalarda, BBT ve ÇAT karşılaştırması yapılan herhangi bir çalışmaya rastlanmamış, bu nedenle ÇAT'ları kendi içerisinde karşılaştıran iki çalışmaya yer verilmiştir.

Doğruöz (2018), ÇAT'ları kendi içerisinde örneklem büyüklüğü, panel deseni ve modül uzunluğu koşullarında, farklı test birleştirme yöntemlerine göre karşılaştırmıştır. Yukarıdan-aşağıya birleştirme yöntemi kullanıldığı durumda, modül uzunluğunun artmasının ve 1-2 panel deseninden 1-2-2 ve 1-2-3 panellerine geçişin

hata değerlerini düşürdüğü sonucuna ulaşılmıştır. Benzer şekilde, aşağıdan yukarıya test birleştirme yöntemi kullanıldığında, modül uzunluğunun ve modül sayısının artışının ortalama hata değerini azalttığı sonucuna ulaşılmıştır. Ayrıca, her iki test birleştirme yönteminde de örneklem büyüklüğünün artışı tüm desenler için yanlılık değerini düşürmüştür.

Boztunç-Öztürk (2019), yönlendirme modülünün uzunluk ve özelliklerinin ölçüm hassasiyetine etkisini farklı panel desenleri altında incelemiştir. Altı farklı modül uzunluğu, dokuz farklı modül özelliği ve iki farklı panel deseni koşullarının ele alındığı bu çalışmada, modül uzunluğu arttıkça RMSE değerinin azaldığı tespit edilmiştir. Modül uzunluğunun 15 ve üzeri olduğu durumda, yönlendirme modülünün a ve c parametre dağılımları literatürde önerilen aralıkların dışında olmasına rağmen, her iki panel deseninde de iyi sonuçlar elde edilmiştir. Ancak, üç aşamalı panel deseni ölçüm hassasiyeti bakımından daha iyi sonuçlar sağlamıştır.

İlgili Araştırmalar Özet

Literatürde bulunan çalışmalar incelendiğinde, BBT yaklaşımları testler üzerinde yapılan pek çok çalışma bulunduğu, özellikle BBT ve ÇAT karşılaştırmalarının farklı koşullar altında birçok kez incelendiği görülmüştür. Ancak, bu testler üzerinde yapılmış DMF çalışmalarının sayısı azdır. Yapılan çalışmalarda ise, DMF'li maddeleri tespit etmede kullanılan yöntemler farklı koşullar altında incelenmiş ve yöntemlerin DMF'li maddeleri doğru tespit edebilme performansları karşılaştırılmıştır. Sadece bir çalışma (Piromsombat, 2014), testte bulunan DMF'li maddelerin testlerin etkililiği üzerindeki etkisini araştırmıştır. Ancak bu çalışmada sadece BBT'ler ele alınmış, ÇAT'lar araştırma kapsamına alınmamıştır. Bu nedenle, testte DMF'li madde bulunmasının ÇAT'ların performansı üzerindeki etkisinin incelendiği ve ÇAT ile BBT performansının karşılaştırıldığı bu çalışmanın literatüre önemli katkılar sağlayacağı düşünülmektedir.

Bölüm 3
Yöntem

Bu bölümde araştırmanın türü, araştırma deseni, verinin üretilmesi, simülasyon süreçleri ve veri analizi ile alakalı detaylı açıklamalara yer verilmiştir.

Araştırmanın Türü

Araştırma kapsamında, testte değişen madde fonksiyonu (DMF) gösteren maddelerin yer almasının, bireyselleştirilmiş bilgisayarlı (BBT) ve çok aşamalı testlerin (ÇAT) etkililiği üzerindeki etkisinin farklı koşullar altında incelenmesi amaçlanmaktadır. Araştırmada kullanılan veriler simülasyon yöntemiyle üretilmiş; farklı test desenleri, farklı koşullar altında kontrollü şekilde karşılaştırılmıştır. Simülasyon çalışmaları bilinen olasılık dağılımlarından sözde-rastgele örnekleme (pseudo-random sampling) yapılarak veri üretimi yapmayı içeren bilgisayar deneyleridir ve belirli senaryolardaki istatistiksel yöntemlerin performansı hakkında deneysel sonuçlar elde etmek için kullanılmaktadır. Psikometri için temel olan simülasyon araştırmaları, özellikle yeni yöntemlerin değerlendirilmesinde ve alternatif yöntemlerin karşılaştırılmasında değerli araçlardır (Feinberg ve Rubright, 2016; Morris, White ve Crowther, 2019). İlgili çalışma, verinin simüle edildiği bir Monte Carlo simülasyon çalışmasıdır. Monte Carlo çalışmaları, birçok yönüyle deneysel çalışmaların görüntülerini yansıtmaktadır; ancak bu çalışmalarda veri bilgisayar aracılığıyla üretilmektedir (Harwell, Stone, Hsu ve Kirisci, 2016). Çalışma kapsamında ele alınan tüm koşulların gerçek veride aynı anda karşılanması zor olduğundan simülasyon verisi tercih edilmiştir. Ayrıca, DMF'li maddelerin BBT ve ÇAT'lar üzerindeki etkisinin ayrıntılı incelenerek ve karşılaştırılarak ortaya çıkarılması yönüyle araştırmanın betimsel araştırma özelliği taşıdığı söylenebilir. Betimsel araştırmalar, verilen bir durumu mümkün olduğunca eksiksiz ve dikkatli bir şekilde tanımlandığı çalışmalardır (Fraenkel, Wallen ve Hyun, 2012).

Araştırma Deseni

Çalışma kapsamında bir BBT, iki ÇAT deseni (1-2-4 ve 1-3-3) farklı koşullar altında karşılaştırılmıştır. Hem BBT hem de ÇAT desenlerinde manipüle edilen ortak koşul test uzunluğu ve DMF'li madde oranıdır. Koşullara ilişkin detaylı bilgi Tablo 1'de sunulmuştur.

Tablo 1

Simülasyon Koşulları

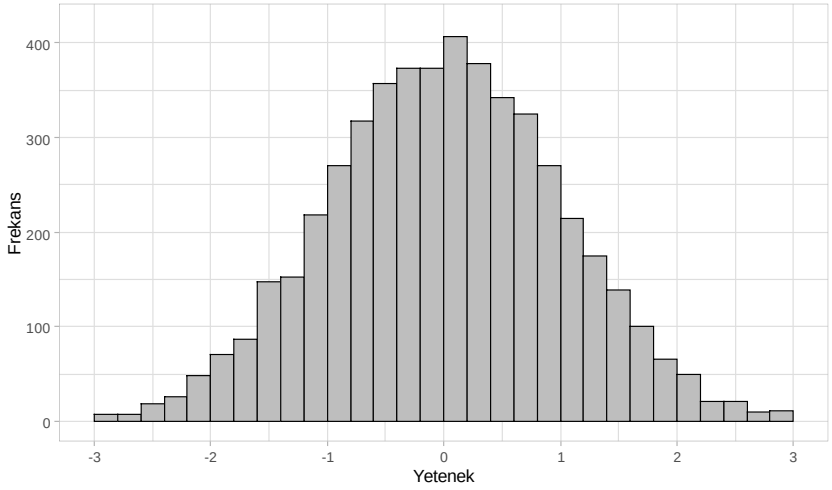
Test Deseni		DMF'li Madde Oranı	Test Uzunluğu
BBT		%10	10 20 30 40
		%20	10 20 30 40
		%30	10 20 30 40
ÇAT	ÇAT 1-3-3	%10	10 20 30 40
		%20	10 20 30 40
		%30	10 20 30 40
	ÇAT 1-2-4	%10	10 20 30 40
		%20	10 20 30 40
		%30	10 20 30 40

Tablo 1'de görüldüğü gibi test desenleri (BBT, ÇAT 1-3-3 ve ÇAT 1-2-4), üç farklı DMF'li madde oranı ve dört farklı test uzunluğu altında incelenmiştir. Tüm koşullar, bulundukları desen içerisinde birbiriyle tamamen çaprazlanmıştır. Örneğin, BBT için toplamda 3 farklı DMF'li madde oranı ve 4 farklı test uzunluğu olmak üzere 12 koşul incelenmiştir. ÇAT deseni içinse bu koşullara ek olarak 2 panel deseni bulunduğundan toplamda 24 koşul ele alınmıştır. Böylelikle, farklı test uzunluklarında ve testteki DMF'li madde oranının farklılaştığı durumlarda test desenlerinin etkililiğinin incelenmesi amaçlanmıştır. Çalışma kapsamında toplam 36 koşul 30 replikasyon ile incelenmiştir.

Verinin Üretilmesi

Verilerin üretilmesi aşamasında ücretsiz ve açık kaynak kodlu istatistiksel bir dil olan R programlama dilinden faydalanılmıştır (R Core Team, 2018). Öncelikle 600 maddelik bir havuz ve 5000 bireye ilişkin yetenek parametresi üretilmiştir. Madde havuzunda bulunan maddelerin bir kısmı araştırmacı tarafından DMF'li hale getirilmiştir. Bu işlemin ardından, BBT ve ÇAT ortamları oluşturulmuş; oluşturulan ortamlar aynı madde havuzu ve aynı bireyler üzerinden farklı koşullar altında karşılaştırılmıştır. Analizler xxIRT (Luo, 2018), catR (Magis, Raiche ve Barrada, 2018) ve mstR (Magis, Yan ve von Davier, 2018) paketleri kullanılarak yapılmıştır. Süreçler ayrıntılı olarak aşağıda açıklanmıştır.

Yetenek parametrelerinin üretilmesi. Çalışma kapsamında, 5000 bireye ilişkin yetenek parametreleri, standart normal dağılım N(0,1) temel alınarak üretilmiştir. Uç değerlerin etkisini ortadan kaldırmak amacıyla yetenek parametreleri [-3,+3] aralığında olacak şekilde sınırlandırılmıştır. Üretilen yetenek parametrelerinin dağılımı Şekil 4'te sunulmuştur.



Şekil 4. Üretilen yetenek parametrelerinin dağılımı

Şekil 4'te görüldüğü gibi, oluşturulan yetenek parametreleri [-3, +3] aralığında kalacak şekilde normal dağılım göstermektedir. Ayrıca yetenek parametrelerine ilişkin betimsel istatistikler incelenmiş; ortalama, standart sapma, çarpıklık ve basıklık değerleri sırasıyla 0.009, 0.996, 0.009, -0.183 olarak hesaplanmıştır. Oluşturulan 5000 kişilik grup rastgele ikiye bölünmüş ve 2500'er kişilik referans ve odak grupları elde edilmiştir.

Madde havuzunun oluşturulması. Madde havuzu oluşturulurken, 3 parametreli lojistik model (3 PLM) temel alınarak 600 maddeye ilişkin parametre üretimi yapılmıştır. Bu aşamada, üç farklı güçlük düzeyi için farklı güçlük parametre dağılımı kullanılmış, ayırıcılık ve şans parametrelerinin dağılımları ise tüm güçlük düzeyleri boyunca aynı parametre dağılımı kullanılarak üretilmiştir. Her bir düzeyde 200'er madde bulunması sağlanmıştır. Madde havuzunun oluşturulmasında temel alınan parametreler Tablo 2'de sunulmuştur.

Tablo 2

Madde Havuzunun Oluşturulmasında Kullanılan Parametre Dağılımları

Güçlük Düzeyi	a parametresi	b parametresi	Madde sayısı	c parametresi
Zor	Uniform [0.5 - 2]	N (1,1)	200	Uniform [0 – 0.25]
Orta		N (0,1)	200	
Kolay		N (-1,1)	200	

Tablo 2'de görüldüğü gibi, tüm zorluk düzeylerinde ayırıcılık parametresi olan a parametresi [0.5, 2] aralığında, şans parametresi olan c parametresi ise [0, 0.25] aralığında Uniform dağılım gösterecek şekilde üretilmiştir. Normal dağılıma uygun olarak üretilen b parametresine ilişkin ortalama değeri, üç farklı güçlük düzeyi boyunca farklılık gösterirken standart sapma değeri sabit tutulmuştur. Kolay, orta ve zor maddeler için ortalama sırasıyla -1, 0 ve 1 olacak şekilde belirlenmiş; standart sapma değeri ise tüm düzeylerde 1 olarak alınmıştır. Ayrıca, güçlük parametrelerinin sadece [-3, +3] aralığında değer alması sağlanmıştır. Baker (2001), madde ayırıcılık parametresinin genellikle 0.5 - 2 aralığında, güçlük parametresinin ise -3 ile +3 arasında değerler aldığını belirtmiştir. Flaugher (2000) ise, bireyselleştirilmiş testler için madde havuzundaki maddelerin a parametresinin 1'den yüksek olmasını önermiştir. Bu bilgilere dayanılarak a parametresi U [0.5, 2.5] dağılımı ile, b parametresi ise yetenek parametrelerinin dağılımıyla uyumlu olması amacıyla normal dağılımdan üretilmiştir. c parametresinin üretiminde ise U [0, 0.25] dağılımı kullanılmıştır. Kullanılan bu dağılımlar ile geniş bir aralıkta ve kullanıma uygun madde parametreleri elde edilmiştir.

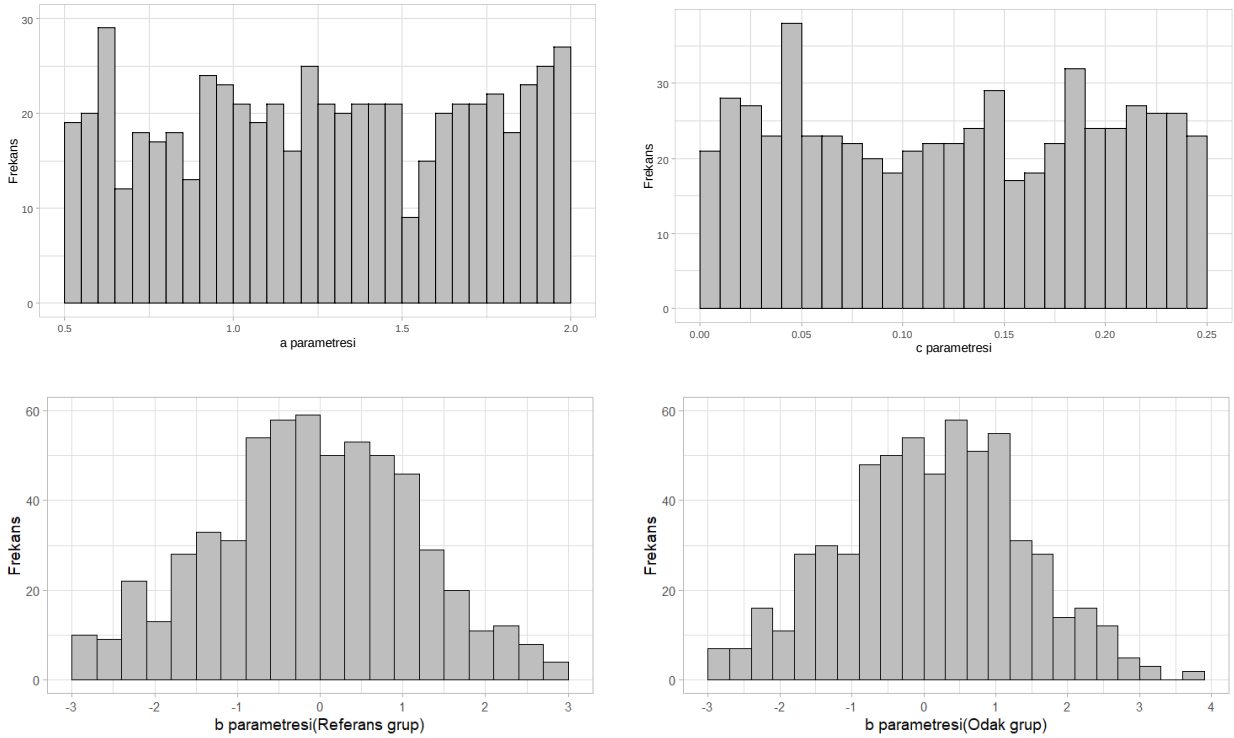
Oluşturulan 600 maddelik havuzdaki maddelere ilişkin betimsel istatistikler Tablo 3'te sunulmuştur.

Tablo 3

Madde Havuzuna İlişkin Betimsel İstatistikler

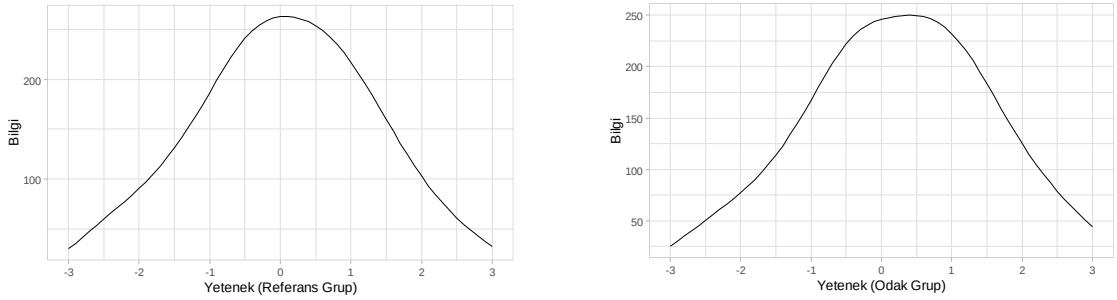
	a parametresi	b parametresi (referans)	b parametresi (odak)	c parametresi
K	600	600	600	600
Ortalama	1.268	-0.097	0.1025	0.125
Standard Sapma	0.442	1.22	1.27	0.074
Minimum	0.501	-2.967	-2.967	0.0002
Maksimum	1.999	2.988	3.870	0.249
Basıklık	-1.208	-0.349	-0.277	-1.303
Çarpıklık	-0.015	-0.061	-0.016	0.003

Tablo 3'te verildiği gibi, a parametresi için ortalama değeri 1.268 standart sapma değeri ise 0.442'dir. Referans ve odak gruplar için ortalama değerleri sırasıyla -0.097 ve 0.103 iken standart sapma değerleri 1.22 ve 1.27'dir. Odak grup için madde havuzunun %20'sinin zorlaştırılmış olması nedeniyle ortalama güçlük değeri yükselmiştir. c parametresi için ortalama ve standart sapma değerleri ise 0.125 ve 0.074'tür. Minimum ve maksimum değerleri tüm parametreler için belirlenen aralıklar içerisinde kalmış, yalnızca odak grup b parametrelerinden beş tanesi +3 sınırının dışına çıkmıştır. Bu durum, madde havuzunun %20'sinin DMF'li hale getirilmesi amacıyla referans grupta +1 değeri eklenmiş olan b parametrelerinden kaynaklıdır. Parametrelere ilişkin dağılımlar Şekil 5'te sunulmuştur.



Şekil 5. Üretilen madde parametrelerinin dağılımı

Elde edilen madde havuzunun test bilgi fonksiyonu Şekil 6'da odak ve referans grup için ayrı ayrı sunulmuştur.

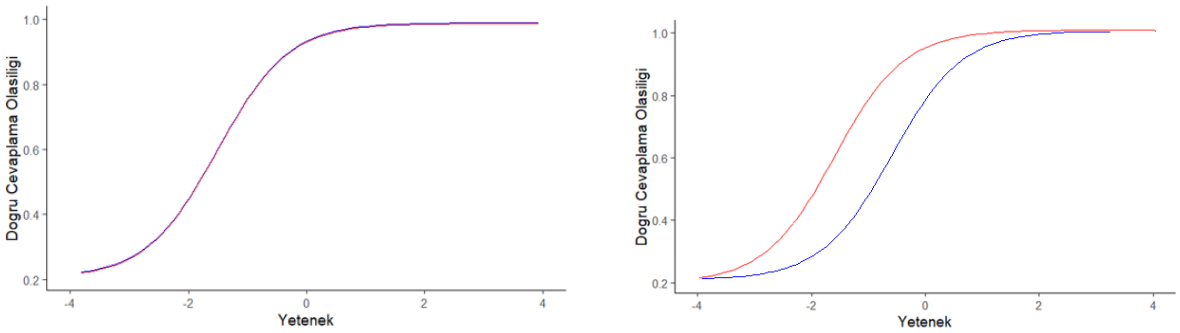


Şekil 6. Madde havuzuna ait test bilgi fonksiyonları

Test bilgi fonksiyonları incelendiğinde, madde havuzunun her iki grup için de yetenek düzeyinin 0 olduğu nokta ve etrafında yüksek bilgi verdiği ve [-3, 3] yetenek aralığını kapsadığı görülmektedir.

Maddelerin DMF'li hale getirilmesi. Madde havuzu, her bir güçlük düzeyinde 200, toplamda 600 madde olacak şekilde oluşturulmuş; daha sonra her düzey için maddelerin %20'si rastgele seçilerek DMF'li hale getirilmiştir. Bu maddeleri DMF'li hale getirmek için referans gruptaki b parametrelerine +1 değeri eklenerek odak grup parametreleri elde edilmiştir. DMF gösteren maddeler için odak ve referans gruplarının b parametreleri arasında +1 fark bulunmakta ($b_{odak} - b_{referans}$

= 1) ve bu maddeler referans grubu lehine çalışmaktadır. Dolayısıyla, elde edilen DMF'li maddelerin tümü tek biçimli DMF (TBDMF) göstermektedir. Sonuç olarak, bir düzeyde 40 TBDMF'li 160 DMF'siz, toplamda ise 120 TBDMF'li 480 DMF'siz madde bulunmaktadır. Şekil 7'de bir maddenin DMF'li hale getirilmeden önceki ve getirildikten sonraki haline ilişkin madde karakteristik eğrileri örnek olarak sunulmuştur.



DMF'siz madde DMF'li madde

Şekil 7. Madde 161'e ait DMF'li hale getirilmeden önceki ve getirildikten sonraki madde karakteristik eğrileri (kırmızı: referans grup, mavi: odak grup)

Şekil 7'de görüldüğü gibi, maddenin DMF'siz halinde her iki grup için de madde karakteristik eğrileri çakışmaktadır. Yani bireylerin maddeyi doğru cevaplama olasılıkları tüm yetenek düzeylerinde eşittir. Madde DMF'li hale getirildiğinde ise düşük yetenek gruplarında referans grubun maddeyi doğru yanıtlama olasılığı odak gruba kıyasla daha yüksektir.

Maddelerin DMF'li hale getirilmesinden sonra kontrol amacıyla DMF analizi yapılmıştır. Raju'nun işaretli alan indeksi yöntemi ile yapılan ilgili analiz için DFIT paketi (Cervantes, 2017) kullanılmıştır. Elde edilen alan ölçümlerinin yorumlanmasında iç ölçütlerden faydalanılabilmektedir. Bu ölçütler, alan ölçümlerine ait ortalama, medyan değerleri ve bu değerlerin bir çeyrek sapma veya standart sapma üstü olabilir (Deveci-Ateşok, 2008; Öğretmen ve Doğan, 2004; Uzun ve Gelbal, 2017). Çalışma kapsamında iç ölçüt olarak maddelerin alan indeksleri değerlerine ait ortalama ve medyan değerinin bir standart sapma üstü kullanılmıştır. Alan ölçümlerine ait ortalama değeri 0.175, medyan değeri 0, standart sapma değeri ise 0.352 olarak hesaplanmıştır. Ortalamanın ve medyanın bir standart sapma üzeri sırasıyla 0.527 ve 0.352'dir. Alan ölçüm değerleri incelendiğinde, yalnızca araştırmacı tarafından DMF'li hale getirilen 120 maddeye ilişkin değerlerin 0'dan

farklılaştığı, kalan 480 maddenin tümüne ilişkin alan ölçümünün 0 olduğu görülmektedir. Bu durum, ilgili 480 maddede DMF görülmediğinin göstergesidir. Diğer 120 maddeye ait alan ölçüm değerleri EK-A'da sunulmuştur. EK-A incelendiğinde, alan ölçümlerinin 0.752 ile 0.999 aralığında değişen değerler aldığı görülmektedir. Bu değerler iç ölçüt değeri olarak kullanılan ortalama ve medyanın bir standart sapma üstü olan 0.527 ve 0.352'nin oldukça üstündedir. Dolayısıyla, araştırmacının DMF'li hale getirdiği 120 maddenin DMF gösterdiği sonucuna ulaşılmıştır.

MTK varsayımlarının incelenmesi. Monte-Carlo simülasyon süreçlerinde tek boyutlu MTK modellerine dayalı işlem yapıldığından, bu modellere ilişkin varsayımlar incelenmiştir. Embretson ve Reise (2000), en yaygın kullanılan MTK modellerinde basit ama güçlü varsayımların bulunduğunu ve iki temel varsayımın tek boyutluluk ve yerel bağımsızlık olduğunu belirtmiştir. Tek boyutluluk, testi oluşturan maddelerin yalnızca bir yeteneği ölçmesi, madde cevaplarındaki ortak varyansı açıklamak için tek bir özelliğin yeterli olması durumudur. Yerel bağımsızlık ise, bireyin yeteneği kontrol edildiğinde, test maddelerinin birbiri ile ilişkili olmamasıdır. Bir diğer deyişle, bireyin herhangi iki maddeye verdiği cevaplar birbirinden bağımsızdır. Bireyin test maddelerine verdiği cevapları etkileyen tek faktör bireyin yeteneğidir (Embretson ve Reise, 2000; Hambleton ve Swaminathan, 1991).

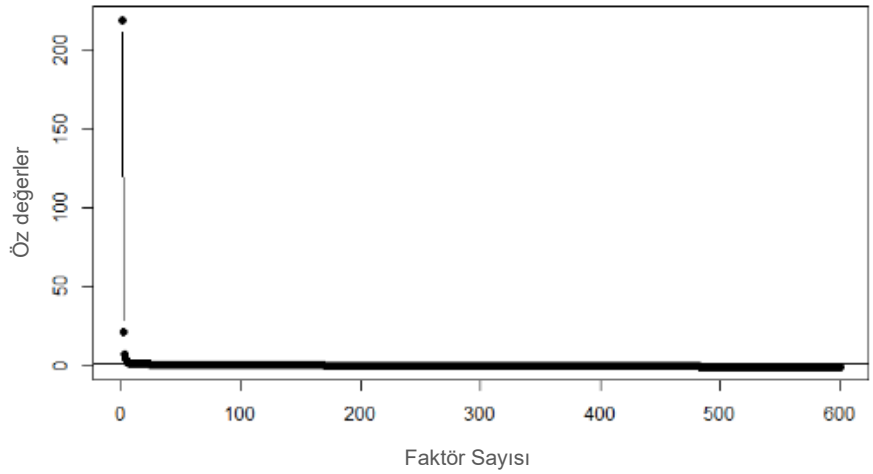
Tek boyutluluk. Tek boyutluluğun incelenmesi amacıyla tetrakorik korelasyon matrisine dayalı açımlayıcı faktör analizi 'psych' paketi (Revelle, 2019) ile yapılmıştır. Faktör analizinden önce, verinin analize uygun olup olmadığını incelemek için Kaiser-Meyer-Olkin (KMO) ve Bartlett küresellik testi değerleri hesaplanmıştır. KMO değerinin 0.60'tan yüksek, Bartlett testinin ise istatistiksel olarak anlamlı çıkması verinin faktör analizine uygun olduğunun göstergesidir (Tabachnick, Fidell ve Ullman, 2007). Hesaplanan KMO değeri (KMO=0.994) ve Bartlett testi (χ^2=375710, sd=599, p=0.00) sonuçlarına göre veri seti faktör analizi için gerekli koşulları sağlamıştır. Ardından, faktör analizi yapılmış ve elde edilen öz değerler incelenmiştir. İlk dört faktöre ilişkin öz değerler Tablo 4'te sunulmuştur.

Tablo 4

Faktör Analizinden Elde Edilen Öz değerler

Faktör Sayısı	Özdeğer
1	219.36
2	22.10
3	8.06
4	4.69

Lord (1980), ilk faktöre ait öz değerin ikinci faktörün öz değerine kıyasla büyük olmasını tek boyutluluğun göstergesi olarak belirtmiştir. Benzer şekilde, Ackerman (1989), ilk öz değerin ikinciye oranının çok boyutluluğa kanıt sağlayabileceğini belirtmiştir. Bu oranın 3 ve üzerinde olması, tek boyutluluğun göstergesi olarak sıkça kullanılan bir kriterdir (Reise, Cook ve Moore, 2014). Öz değer tablosunda görüldüğü gibi, birinci faktöre ilişkin öz değer ikinci faktöre ilişkin öz değerin 9 katından fazladır. Ayrıca, Şekil 8'de verilen yamaç-birikinti grafiği oluşturulmuş ve incelenmiştir.



Şekil 8. Yamaç-Birikinti grafiği

Grafik incelendiğinde, birinci faktöre ait öz değerden sonra oldukça hızlı bir düşüş olduğu görülmektedir. Bu durum tek boyutluluğun göstergesi olarak yorumlanabilir. Son olarak, tek faktör tarafından açıklanan varyans oranı incelenmiş, tek faktörün toplam varyansın %36.4'ünü açıklamakta olduğu görülmüştür. Reckase (1979), açıklanan varyansın %20 ve üzerinde olmasının tek boyutluluk için yeterli olduğunu belirtmiştir. Sonuç olarak tüm bilgiler birlikte ele alındığında, verinin tek boyutluluk varsayımını sağladığı görülmektedir.

Yerel bağımsızlık. Tek boyutluluk varsayımının sağlanması, yerel bağımsızlık için bir kanıt olarak kabul edilmektedir. Yerel bağımsızlık varsayımının sağlanmış olması; bireyin herhangi iki maddeye verdiği cevabın birbirini etkilememesi, cevapların yalnızca bireyin yeteneğinden etkilenmesi anlamına gelmektedir. Bu nedenle test tek boyutlu ise yerel bağımsızlık varsayımının da sağlandığı söylenebilir (Hambleton ve Swaminathan, 1991; Embretson ve Reise, 2000). Tek boyutluluk varsayımının sağlandığı yukarıda gösterilmiştir. Bu durum yerel bağımsızlığın da sağlandığının bir göstergesi olarak kabul edilebilir. Ek olarak yerel bağımsızlığın incelenmesinde Yen'in Q_3 değerlerine 'subscore' paketi (Dai, Wang ve Svetina, 2019) ile bakılmıştır. Yerel bağımsızlığın göstergesi olan Q_3 istatistiği test performansını kontrol altına aldıktan sonra iki maddeye verilen cevaplar arasındaki korelasyonu incelemektedir (Yen, 1984). Chen ve Thissen (1997), 0.20'den yüksek değerlerin yerel bağımsızlığın ihlaline işaret ettiğini belirtmiştir. Analiz sonucu elde edilen Q_3 istatistiklerinin hiçbiri kritik değer olan 0.20'yi aşmamıştır. Sonuç olarak, yerel bağımsızlık varsayımının sağlanmış olduğu sonucuna ulaşılmıştır.

BBT ve ÇAT simülasyonları. Oluşturulan BBT ve ÇAT simülasyonları için sırasıyla catR ve xxIRT-mstR paketleri kullanılmıştır. Aynı madde havuzu ve aynı bireyler üzerinden yürütülmüş olan bu simülasyonlar için ortak olarak manipüle edilen değişken, test uzunluğu (10-20-30-40) ve testteki DMF gösteren madde oranı (%10 - %20 - %30) dır. Ayrıca, ÇAT uygulamasının kendi içerisinde panel desenleri (1-3-3, 1-2-4) manipüle edilmiştir. Dolayısıyla, 12 farklı BBT (4 test uzunluğu x 3 DMF oranı) ve 24 farklı ÇAT (4 test uzunluğu x 3 DMF oranı x 2 panel deseni) koşulu ortaya çıkmıştır. Her bir koşul için 30 replikasyon yapılmıştır. Yapılan 30 replikasyondan elde edilen RMSE, yanlılık ve korelasyon değerlerinin ortalaması alınmış, elde edilen değer nihai değer olarak yorumlanmıştır. Daha iyi karşılaştırma yapılabilmesi amacıyla, BBT ve ÇAT uygulamalarında; maksimum madde kullanım oranı, MTK modeli, yetenek kestirim yöntemi ve madde/modül seçim yöntemi sabit tutulmuştur. Maksimum madde kullanım oranı BBT için 0.25 olarak sabitlenmiş, ÇAT içinse dört ayrı paralel panel oluşturularak maksimum madde kullanım oranının 0.25 olması sağlanmıştır. Her iki uygulama için de bireylerin cevapları 3 PLM'ye göre üretilmiştir. Yetenek kestirimi için EAP yöntemi (önsel dağılım N(0,1)), madde/modül seçim yöntemi olarak Maksimum Fisher Bilgisi yöntemi seçilmiştir. Böylelikle, iki

uygulama da benzer şekilde uygulanabilmiştir. Bütün simülasyon süreçleri birbirinin eşleniği olan catR ve mstR paketleri yardımıyla gerçekleştirilmiştir. BBT ve ÇAT uygulamalarına ilişkin detaylı açıklamalar aşağıda verilmiştir.

***BBT simülasyonu*.** BBT ortamı, catR paketi aracılığıyla oluşturulmuş; simülasyonda dört farklı test uzunluğu (10-20-30-40) ve üç farklı DMF'li madde oranı (%10 - %20 - %30) olmak üzere toplam 12 farklı koşul (4 test uzunluğu x 3 DMF'li madde oranı) incelenmiştir. Madde seçiminde Maksimum Fisher Bilgisi yöntemi, yetenek kestiriminde ise EAP kestirim yöntemi kullanılmıştır. Başlama kuralı olarak, başlangıç yetenek düzeyi 0 olarak belirlenmiş ve her koşul için bu değer kullanılmıştır. Bu kurala göre, bireylerin başlangıçtaki yetenek düzeyleri '0' (sıfır) kabul edilmiş ve bireyin karşılaşacağı ilk madde buna göre belirlenmiştir.

***ÇAT simülasyonu*.** ÇAT ortamının oluşturulmasında, xxIRT ve mstR paketlerinden faydalanılmıştır. ÇAT simülasyonunda iki farklı panel deseni (1-3-3 ve 1-2-4), dört farklı test uzunluğu (10-20-30-40) ve üç farklı DMF'li madde oranı (%10 - %20- %30) olmak üzere toplam 24 koşul (4 test uzunluğu x 3 DMF'li madde oranı x 2 panel deseni) incelenmiştir. Madde seçim yöntemi olarak BBT'de olduğu gibi Maksimum Fisher Bilgisi, yetenek kestirim yöntemi olarak ise EAP yöntemi tercih edilmiştir.

Oluşturulan 1-3-3 ÇAT ortamında başlangıç aşamasında tek bir modül kullanılırken, ikinci ve üçüncü aşamalarda üçer modül yer almaktadır. Toplam 7 modülün yer aldığı bu desende, ilk aşamada tüm bireyler için ortak olan tek bir modül oluşturulmuş ve bu modülün zorluk düzeyi orta düzey olarak belirlenmiştir. İkinci ve üçüncü aşamalarda yer alan üçer modül ise üç farklı zorluk düzeyine (kolay, orta, zor) sahiptir. Her birey toplamda üç modül cevaplamıştır. Benzer şekilde 1-2-4 panel deseninde de bireyler toplam üç modüle yanıt vermişlerdir. İlk aşamada bulunan tek bir modülü cevaplayan bireyler, buradan elde edilen yetenek kestirimlerine göre ikinci aşamadaki iki farklı düzeydeki modülden birine yönlendirilmiştir. Bu aşamayı da tamamlayan bireyler, ikinci aşama sonucunda kestirilen yetenekleri dikkate alınarak üçüncü aşamadaki dört modülden birine yönlendirilmiştir. Modüllerde yer alan madde sayıları ve bir panelin oluşturulması için gerekli madde sayıları test uzunluklarına ve panel desenine göre farklılık göstermiştir. Modüllerdeki madde sayıları ve panel için kullanılan madde sayıları Tablo 5'te detaylı şekilde sunulmuştur.

Tablo 5

Modüllerdeki ve Panellerdeki Madde Sayıları

Panel Deseni		Test Uzunluğu			
		10	20	30	40
1-3-3	Modül Uzunlukları	3-3-4	6-7-7	10-10-10	13-13-14
	Panelde Kullanılan Madde Sayısı	24	48	70	94
1-2-4	Modül Uzunlukları	3-3-4	6-7-7	10-10-10	13-13-14
	Panelde Kullanılan Madde Sayısı	25	48	70	95

Bir panelde farklı düzeylerde modüller yer aldığı için, ilgili panel oluşturulurken kullanılan madde sayısı, test uzunluğundan daha fazladır. Örneğin; 1-3-3 deseninde, test uzunluğunun 40 olduğu koşulda, bireyler ilk iki aşamada 13'er son aşamada 14 olmak üzere toplam 40 madde yanıtlamışlardır. Ancak; ilk aşamada 13 maddeye, ikinci ve üçüncü aşamalarda bulunan üç farklı düzeydeki modüllerde sırasıyla 39 ve 42 maddeye ihtiyaç duyulmuş ve toplamda 94 madde kullanılmıştır.

Her iki panel deseni için de ikinci aşamadaki modüllerdeki maddelerin %10, %20 ve %30'unun DMF'li maddeler arasından seçilmesi sağlanmıştır. Örneğin; test uzunluğunun 10 olduğu durumda, DMF'li madde oranının %20 olduğu koşul altında, maddelerin 8'i DMF göstermeyen, 2'si DMF gösteren maddeler arasından seçilmiştir. Seçilen 2 DMF'li maddenin ikinci aşamadaki modüllerde yer alması sağlanmıştır.

Çalışma kapsamında dört farklı panel oluşturulmuş, böylelikle maksimum panel, modül ve madde kullanım sıklığının BBT ile karşılaştırılabilir hale gelmesi sağlanmıştır. Panellerin oluşturulması süreci aşağıda detaylı bir şekilde açıklanmıştır.

Panellerin oluşturulması. Oluşturulan dört farklı panel xxIRT paketinde bulunan açık kaynak kodlu bir "karışık tamsayılı doğrusal programlama çözücü" (mixed integer linear programming solver) (lp_solve 5.5) aracılığıyla elde edilmiş,

bir panelde kullanılan maddenin bir diğer panelde yer almaması sağlanmıştır. Panellerin oluşturulmasında "Aşağıdan yukarıya test birleştirme" (Bottom-up) yöntemi kullanılmıştır. Bu yöntemde öncelikle her bir modül için dört farklı paralel form oluşturulmuştur. Modüllerin paralel olmasının sağlanabilmesi amacıyla, modül düzeyinde bilgi fonksiyonu hedefleri belirlenmiş ve modüller bu hedefleri sağlayacak şekilde yapılandırılmıştır. Modüllerdeki maddeler, belirtilen yetenek düzeylerinde maksimum bilgi sağlayacak şekilde seçilmiştir. Modül bilgilerinin maksimum hale getirildiği noktalar Tablo 6'da sunulmuştur.

Tablo 6

Modül Bilgilerinin Maksimum Hale Getirildiği Yetenek Noktaları

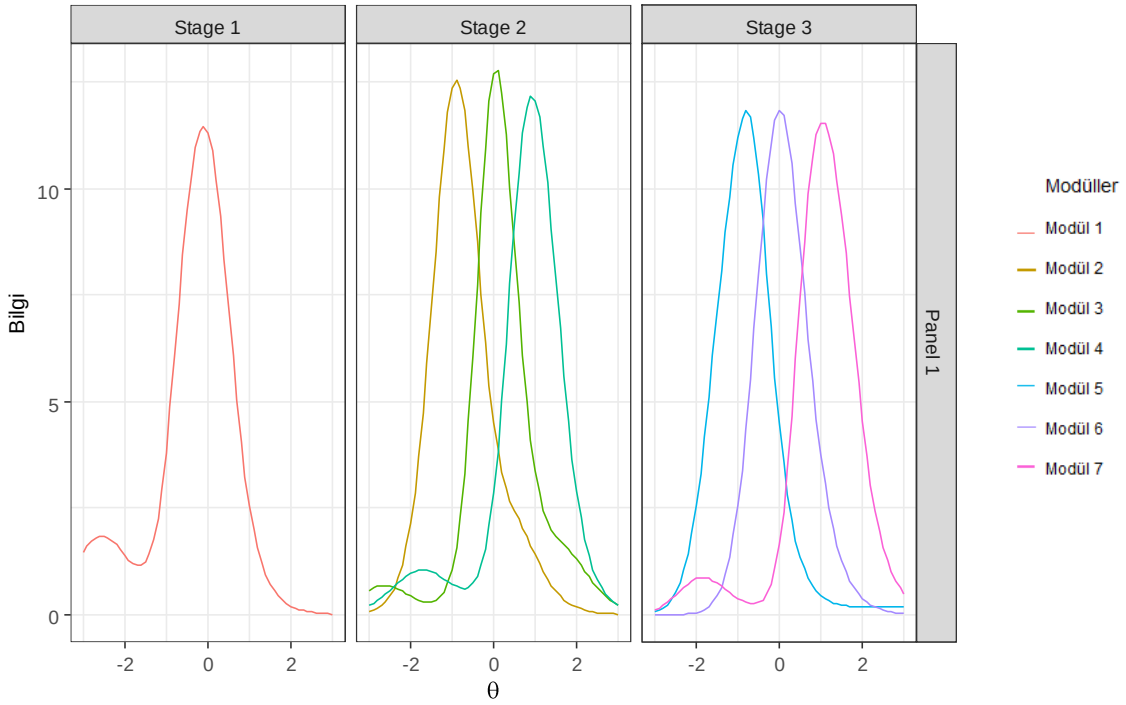
Panel Deseni	Aşama 1	Aşama 2	Aşama 3
1-3-3	θ = 0	θ = (-1,0,1)	θ = (-1,0,1)
1-2-4	θ = 0	θ = (-1,1)	θ = (-1.5, -0.5, 0.5, 1.5)

Tablo 5'te görüldüğü gibi, her iki panel deseninde de ilk aşamada birer modül bulunduğu için, test bilgi fonksiyonu için tek bir merkez (θ= 0) belirlenmiştir. Fonksiyon bu noktada maksimum değerini almaktadır. Diğer yandan ikinci aşamada 1-3-3 deseni için üç farklı modül bulunması nedeniyle, üç farklı merkezli bilgi fonksiyonu (θ = (-1,0,1)) belirlenmiştir. Belirlenen bu noktalar kolay, orta ve zor modülleri birbirinden ayırmıştır. Kolay modüller için -1, orta zorluk düzeyindeki modüller için 0, zor modüller için 1 yetenek düzeyi ve çevresinde modül bilgi fonksiyonunun maksimum değere ulaşması sağlanmıştır. 1-2-4 deseninde ise, ikinci aşamada iki farklı modül bulunduğu için, kolay modül için -1 ve zor modül için +1 yetenek düzeyi ve çevresinde modül bilgi fonksiyonu tepe noktası oluşturacak şekilde iki farklı değer belirlenmiştir. Aşama üç içinde yorumlar benzer şekilde yapılabilir.

Her bir modül için dört paralel formun oluşturulmasından sonra, bu paralel formlar panellere rastgele atanarak paralel paneller elde edilmiştir. Oluşturulan modüllerin paralel olması sayesinde, bu modüller paneller arasında dönüşümlü olarak kullanılabilmektedir (Yan, von-Davier ve Lewis, 2014). Örneğin; 1-3-3 panel deseninde, 1-2K-3O rotasını alan bir bireye ilk aşamada dört paralel modülden biri, ikinci aşamada kolay düzeydeki dört modülden biri, üçüncü aşamada ise orta

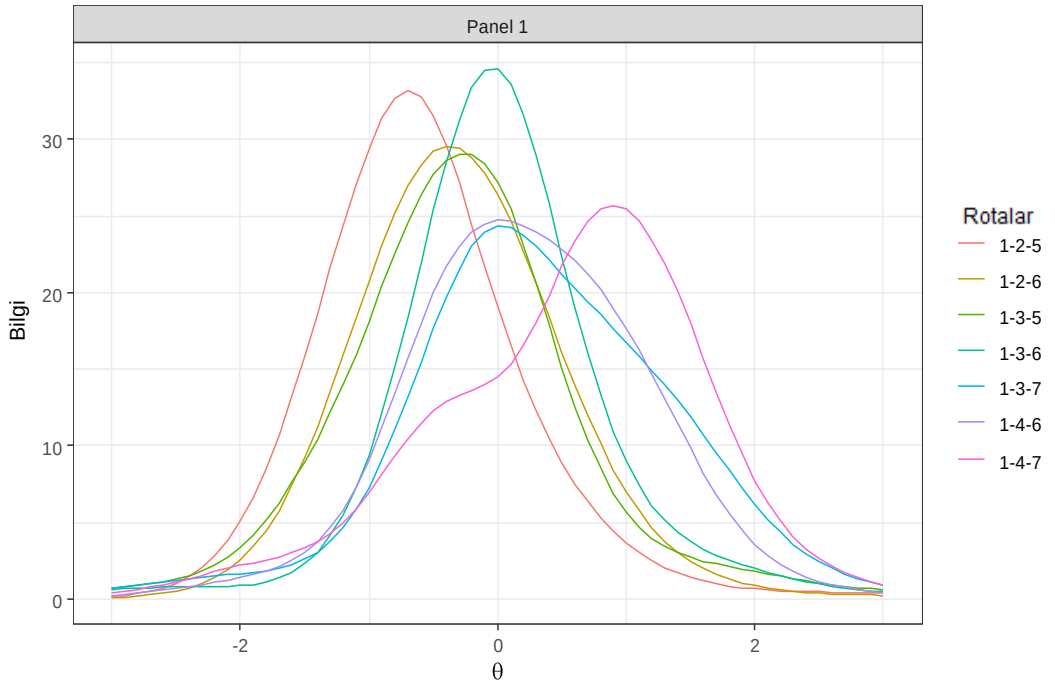
güçlükteki dört paralel modülden biri gelebilmekte ve bu rota için toplamda 64 farklı paralel panel oluşturulabilmektedir.

Şekil 8 ve 9'da örnek birer modül ve rota bilgi fonksiyonu sunulmuştur.



Şekil 8. Örnek modül bilgi fonksiyonu (1-3-3 panel deseni_30 madde)

Şekil 8'de verilen modül bilgi fonksiyonu, 30 maddelik 1-3-3 desenindeki bir ÇAT uygulamasında kullanılan bir panele aittir. Grafikler incelendiğinde, üç farklı aşamadaki (Stage 1, Stage 2, Stage 3) orta zorluktaki modüllerin (Modül 1, 3 ve 6) yetenek düzeyinin 0 (sıfır) olduğu nokta ve etrafında maksimum bilgiyi sağladığı görülmüştür. Kolay modüller (Modül 2 ve 5) yetenek düzeyinin -1 olduğu nokta ve etrafında, zor modüller ise (Modül 4 ve 7) yetenek düzeyinin 1 olduğu noktalar etrafında bulundukları aşamalarda maksimum bilgiyi sağlamıştır.



Şekil 9. Örnek rota bilgi fonksiyonu (1-3-3 panel deseni_30 madde)

Şekil 9'da verilen örnek rota bilgi fonksiyonu da 30 maddelik 1-3-3 desenindeki MST uygulamasında kullanılan bir panele aittir. Rota bilgi fonksiyonu, bu paneli alan bireyin izlediği rotanın sağladığı bilgi miktarını göstermektedir. Örneğin; 1-4-7 rotasında, bireye ilk aşamada orta güçlük düzeyinde, ikinci ve üçüncü aşamalarda ise zor modüller uygulanmıştır. Görüldüğü üzere bu rota yetenek düzeyinin yüksek olduğu noktalar etrafında daha yüksek bilgi sağlamıştır.

Verilerin Analizi

Verilerin analizinde, BBT ve ÇAT'lardan elde edilen sonuçların değerlendirilmesi için Hata Kareleri Ortalamasının Karekökü (Root Mean Square Error – RMSE), yanlılık ve kestirilen ve gerçek yetenek parametreleri arasındaki korelasyon (ρ) değerleri kullanılmıştır. $\hat{\theta}_j$ kestirilen yetenek parametresini, θ_j gerçek yetenek parametresini, N ise toplam birey sayısını temsil etmek üzere; RMSE ve yanlılık değerleri aşağıdaki formüller yardımıyla hesaplanmıştır.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}}$$

$$Yanlılık = \frac{\sum_{j=1}^{N}\left|(\hat{\theta}_j - \theta_j)\right|}{N}$$

Korelasyon değeri ise, $\sigma_{\hat{\theta}_j}$ ve σ_{θ_j} sırasıyla kestirilen ve gerçek yetenek parametrelerine ilişkin standart hata değerleri olmak üzere, aşağıdaki formül aracılığıyla elde edilmiştir.

$$\rho_{\hat{\theta}_j,\theta_j} = \frac{cov\,(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j}\sigma_{\theta_j}}$$

Yapılan 30 replikasyonun her biri için RMSE, yanlılık ve korelasyon değerleri hesaplanmış; bu değerlerin ortalaması alınarak yorumlanmıştır. Hesaplanan değerlerden yola çıkılarak, oluşturulan iki ÇAT ve bir BBT uygulamasından hangisinin diğerlerine göre daha yüksek ölçüm hassasiyeti verdiği farklı koşullar altında değerlendirilmiştir.

Yapılan betimsel değerlendirmelerden sonra, test desenleri (BBT, ÇAT 1-3-3 ve ÇAT 1-2-4) arasındaki farkların manidar düzeye ulaşıp ulaşmadığı ANOVA analizleriyle incelenmiştir. RMSE, yanlılık ve korelasyon değerlerinin ayrı ayrı bağımlı değişken, test deseninin ise bağımsız değişken olarak alındığı üç ayrı tek yönlü ANOVA analizi her bir alt problem için yapılmıştır. Aralarında manidar düzeyde farklılık ortaya çıkan desen grupları için Post-Hoc analizleri yapılmış ve sonuçlar yorumlanmıştır.

Bölüm 4
Bulgular ve Yorumlar

BBT ve ÇAT uygulamalarından elde edilen sonuçlar bu bölümde sunulmuş, her bir alt probleme yönelik elde edilen bulgulara alt başlıklarda yer verilmiştir.

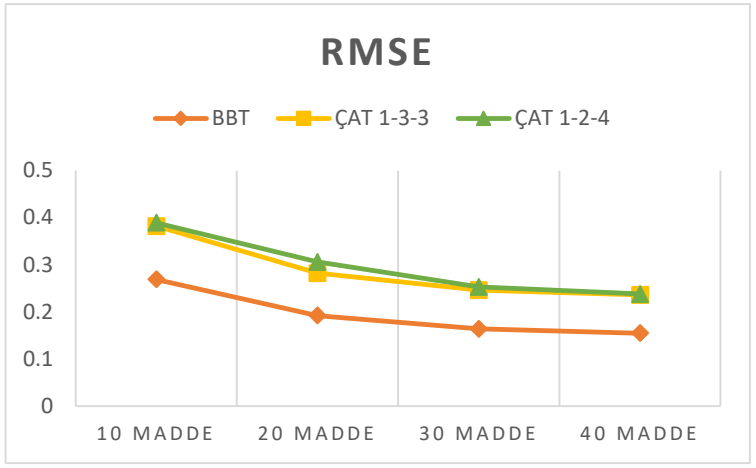
Alt Problem 1'e İlişkin Bulgu ve Yorumlar

"Testte bulunan DMF'li madde oranı %10 iken, farklı test uygulamalarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE, yanlılık ve korelasyon değerleri, test uzunluğuna göre nasıl değişmektedir?"

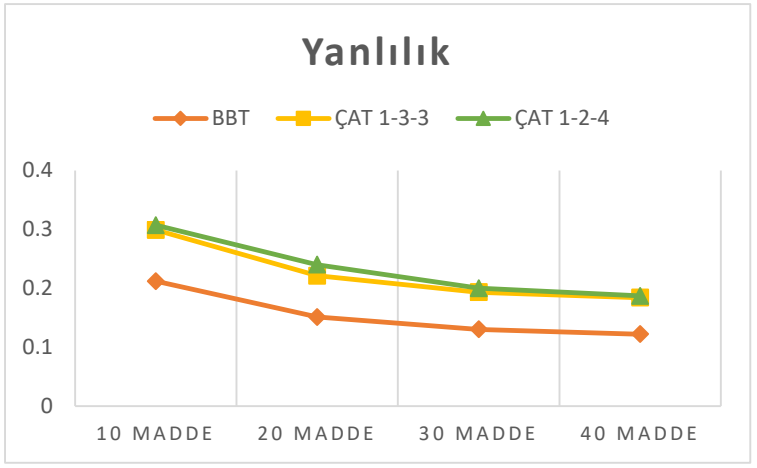
İlgili alt problemde, testte bulunan DMF'li madde oranı %10 olarak sabit tutulmuş ve üç farklı test deseninin etkililiği farklı test uzunlukları altında incelenmiştir. Test desenlerinin etkililiğinin incelenmesinde 30 replikasyondan elde edilen RMSE, yanlılık ve korelasyon değerleri ortalamalarından yararlanılmıştır. Üç farklı test desenine ilişkin RMSE, yanlılık ve korelasyon grafikleri Şekil 10'da sunulmuş ve grafikler üzerinde önemli görülen noktalar yorumlanmıştır. Bu değerlere ilişkin detaylar ise EK-B'de sunulmuştur.

RMSE değerleri BBT deseni için [0.155, 0.269], ÇAT 1-3-3 deseni için [0.236, 0.382] ve ÇAT 1-2-4 deseni için [0.238, 0.389] aralıklarında değişmektedir (EK-B). Şekil 10a'da verilen RMSE grafiğine bakıldığında tüm test uzunlukları için BBT uygulaması en düşük, ÇAT 1-2-4 uygulaması ise en yüksek RMSE değerine sahiptir. Ancak ÇAT 1-3-3 ve 1-2-4 desenlerinde tüm test uzunlukları boyunca RMSE değerleri birbirine oldukça yakın görünmektedir. Ayrıca madde sayısı arttıkça RMSE değeri tüm desenler için azalmaktadır.

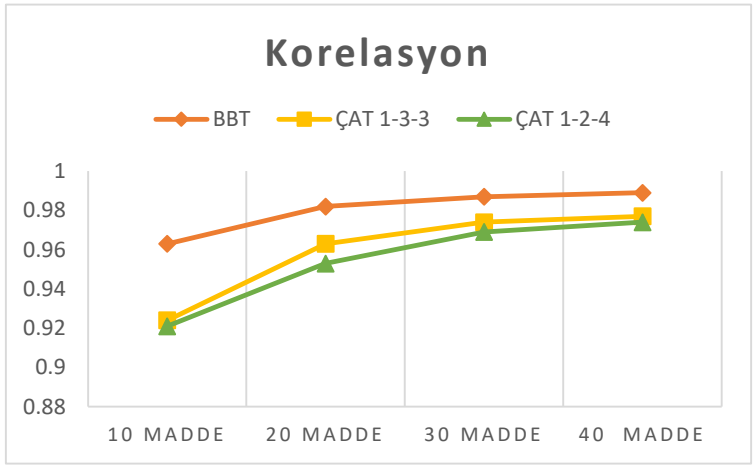
RMSE'ye benzer olarak yanlılık noktasında da tüm test uzunlukları için BBT uygulaması en düşük değere sahiptir. ÇAT 1-3-3 ve ÇAT 1-2-4 desenlerinin ise yakın yanlılık değerlerine sahip olduğu söylenebilir ancak yine de 1-2-4 deseni daha yüksek yanlılık değerlerine sahiptir. Ayrıca, madde sayısı arttıkça yanlılık değerlerinin düştüğü ve desenler arasındaki farkın azaldığı gözlenmiştir. Yanlılık değerleri BBT deseni için [0.122 - 0.212], ÇAT 1-3-3 deseni için [0.184-0.299] ve ÇAT 1-2-4 deseni için ise [0.187-0.307] aralığında değişmiştir (EK-B).



Şekil 10a. DMF oranı %10 iken RMSE değerleri



Şekil 10b. DMF oranı %10 iken yanlılık değerleri

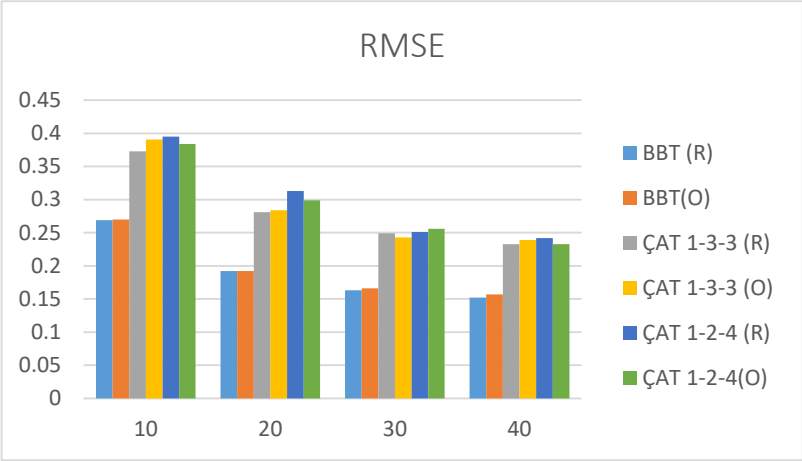


Şekil 10c. DMF oranı %10 iken korelasyon değerleri

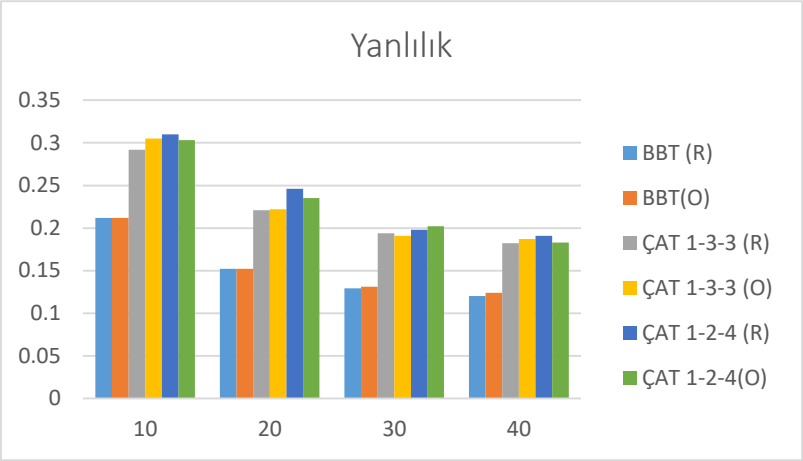
Şekil 10. RMSE, yanlılık ve korelasyon grafikleri (DMF oranı %10)

Son olarak korelasyon değerleri incelendiğinde, BBT için [0.963-0.989] aralığında, ÇAT 1-3-3 deseni için [0.924-0.977] aralığında, ÇAT 1-2-4 için ise [0.921-0.974] aralığında değiştiği görülmektedir (EK-B). Şekil 10c'deki korelasyon grafiğine bakıldığında, tüm test uzunlukları boyunca en yüksek korelasyon değerine sahip olan desenin BBT, en düşük korelasyon değerine sahip desenin ise ÇAT 1-2-4 deseni olduğu saptanmıştır. Tüm test desenleri için madde sayısı arttıkça korelasyon değerleri de artmıştır. Ayrıca, Şekil 10c'de görülebileceği gibi, test uzunluğu arttıkça desenlere ilişkin korelasyon değerleri de birbirine yaklaşmıştır.

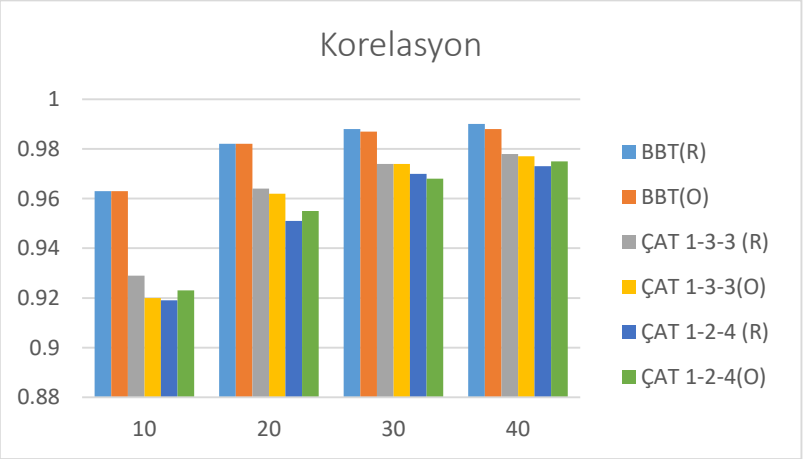
Şekil 10'da verilen grafiklerden ayrı olarak; RMSE, yanlılık ve korelasyon değerlerinin odak ve referans gruplarına göre değişimlerinin incelenmesi amacıyla daha detaylı grafikler çizilmiş ve Şekil 11'de sunulmuştur. Şekil 11a'daki RMSE grafiği incelendiğinde, BBT için 10 ve 20 maddelik testlerde grup değerlerinin birbirine oldukça yakın olduğu, 30 ve 40 maddelik testlerde ise gruplar arası farkın arttığı gözlenmiştir. Bu testlerde odak gruba ilişkin RMSE değeri daha yüksektir. ÇAT desenlerinde ortaya çıkan gruplar arası fark BBT desenindekinden daha fazladır. ÇAT 1-3-3 deseni için 10, 20 ve 40 maddelik testlerde odak grup, 30 maddelik testte ise referans grup daha yüksek hata değerine sahiptir. ÇAT 1-2-4 deseninde ise, 1-3-3 deseninin aksine 10, 20 ve 40 maddelik testlerde referans grupta 30 maddelik testte ise odak grupta daha yüksek hata değeri gözlenmektedir. Şekil 11b'de verilen yanlılık grafiği, RMSE grafiğiyle oldukça benzer olduğu için aynı doğrultuda yorumlar yanlılık için de yapılabilir. Şekil 11c'de verilen korelasyon grafiği incelendiğinde, BBT için 10 ve 20 maddelik testlerde grupların oldukça benzer korelasyon değerlerine sahip olduğu yorumu yapılabilir. 30 ve 40 maddelik testlerde gruplar arası fark artmış, referans gruba ait değer odak grubun değerinden daha yüksek hale gelmiştir. ÇAT 1-3-3 deseninde 30 maddelik testte oldukça yakın olan grup korelasyon değerleri, diğer test uzunluklarında referans grupta daha yüksektir. ÇAT 1-2-4 deseninde ise 30 maddelik testte referans grupta, diğer test uzunluklarında ise odak grupta daha yüksek korelasyon değerleri gözlenmiştir.



Şekil 11a. Gruplara göre RMSE değerleri



Şekil 11b. Gruplara göre yanlılık değerleri



Şekil 11c. Gruplara göre korelasyon değerleri

Şekil 11. Odak ve referans gruplarına göre RMSE, yanlılık ve korelasyon değerleri (DMF oranı %10)

Grafiksel yorumlamalar sonrasında RMSE, yanlılık ve korelasyon değerlerinin test desenleri arasında manidar düzeyde farklılaşıp farklılaşmadığını gözlemlemek için ayrı ayrı tek yönlü ANOVA testleri yapılmıştır. Ancak, korelasyon değerleri normal dağılım göstermediği için (Silver ve Dunlop, 1987), bu değerlerin doğrudan ANOVA analizine alınması uygun değildir. Bu durumda korelasyon değerlerine Fisher'in z dönüşümü uygulanması ve normale yakın dağılım gösteren değerler elde edilmesi önerilmektedir. Fizher'in z dönüşümü korelasyon değerleri dağılımını normalize eder ve böylece dağılımın çarpık olmasından daha az etkilenen bir ortalama korelasyon değerinin elde edilmesinde kullanılabilir (Corey, Dunlap ve Burke, 1998; Silver ve Dunlop, 1987). Bu nedenle, korelasyon değerleri ANOVA analizine doğrudan sokulmamış, Fisher'in z dönüşümü uygulanarak elde edilen değerler analize alınmıştır. RMSE ve yanlılık değerleri için böyle bir durum söz konusu değildir.

Test öncesinde normal dağılım ve varyansların homojenliği varsayımları incelenmiştir. Normal dağılımın kontrolü için normallik testleri ve histogramlara bakılmış, varyansların homojenliği içinse Levene testi sonucu incelenmiştir. Yapılan analizlere ilişkin detaylı bilgiler her bir bağımlı değişken özelinde aşağıda sunulmuştur.

RMSE, yanlılık ve korelasyon değerlerinin bağımlı değişken, test deseninin bağımsız değişken olarak alındığı üç ayrı tek yönlü ANOVA analizi yapılmış, elde edilen bulgular her bir test uzunluğu özelinde ayrı ayrı incelenmiştir. Analizlerde normal dağılım varsayımı sağlanırken, varyansların homojenliği varsayımı bazı durumlarda ihlal edilmiştir. Varsayımın ihlal edildiği durumlarda Welch testinden yararlanılmış, diğer durumlarda ANOVA tablosundaki veriler yorumlanmıştır. Üç analize ilişkin sonuçlar Tablo 7'de sunulmuştur.

Tablo 7

RMSE, Yanlılık ve Korelasyon için Tek Yönlü ANOVA Sonuçları (DMF oranı %10)

		Test Uzunluğu			
Test Deseni		10	20	30	40
RMSE	Test İstatistiği	10154.842 (F)	22123.741 (W)	26989.921 (W)	32389.082 (W)
	p değeri	.000	.000	.000	.000
Yanlılık	Test İstatistiği	24423.021 (W)	69165.800 (W)	17212.486 (W)	3594.671 (W)
	p değeri	.000	.000	.000	.000
Korelasyon	Test İstatistiği	11845.330 (W)	15829.182 (F)	21250.481 (F)	21931.792 (F)
	p değeri	.000	.000	.000	.000

* (F) simgesi varyansların homojenliğinin sağlandığını ve ANOVA F istatistiğinin yorumlandığını
(W) simgesi varyansların homojenliğinin sağlanmadığını ve Welch istatistiğinin yorumlandığını göstermektedir.

Tablo 7'de görüldüğü üzere, RMSE, yanlılık ve korelasyon değişkenlerinin tümü için her bir test uzunluğunda test desenleri arasında manidar düzeyde farklılık göstermektedir ($p<.05$). Farklılıkların hangi gruplardan kaynaklandığını tespit etmek için yapılan Post-Hoc karşılaştırması, homojenliğin sağlandığı gruplarda Tukey, sağlanmadığı gruplarda Dunnett C testi ile yapılmış, alınan sonuçlar Tablo 8'de sunulmuştur.

Tablo 8'de verilen Post-Hoc karşılaştırması sonuçlarına göre, RMSE değerleri farkı tüm test uzunluklarında, tüm desenler arasında manidar düzeye ulaşmıştır. Her bir test uzunluğu için; BBT, ÇAT 1-3-3 ve ÇAT 1-2-4 desenlerine ilişkin RMSE değerlerinin manidar düzeyde farklılaşmakta olduğu sonucuna ulaşılmıştır. Yanlılık ve korelasyon değerleri için de aynı durum söz konusudur. Hem yanlılık hem de korelasyon farkları tüm test uzunluklarında tüm desenler için manidar düzeye ulaşmıştır.

Tablo 8

RMSE, Yanlılık ve Korelasyon için Post-Hoc Sonuçları (DMF oranı %10)

Test Deseni (I)	Test Deseni (J)	Ortalama Farkı (I-J)											
		RMSE				Yanlılık				Korelasyon (z)			
		10 (T)	20 (D)	30 (D)	40 (D)	10 (D)	20 (D)	30 (D)	40 (D)	10 (D)	20 (T)	30 (T)	40 (T)
BBT	ÇAT 1-3-3	-.112*	-.090*	-.082*	-.081*	-.087*	-.069*	-.063*	-.062*	.368*	.365*	.361*	.362*
BBT	ÇAT 1-2-4	-.120*	-.114*	-.089*	-.083*	-.095*	-.088*	-.070*	-.065*	.387*	.479*	.444*	.425*
ÇAT 1-3-3	BBT	.112*	.090*	.082*	.081*	.087*	.069*	.063*	.062*	-.368*	-.365*	-.361*	-.362*
ÇAT 1-3-3	ÇAT 1-2-4	-.008*	-.024*	-.007*	-.002*	-.008*	-.019*	-.007*	-.003*	.019*	.114*	.084*	.063*
ÇAT 1-2-4	BBT	.120*	.114*	.089*	.083*	.095*	.088*	.070*	.065*	-.387*	-.479*	-.444*	-.425*
ÇAT 1-2-4	ÇAT 1-3-3	.008*	.024*	.007*	.002*	.008*	.019*	.007*	.003*	-.019*	-.114*	-.084*	-.063*

*. Ortalama Farkı .05 düzeyinde anlamlıdır

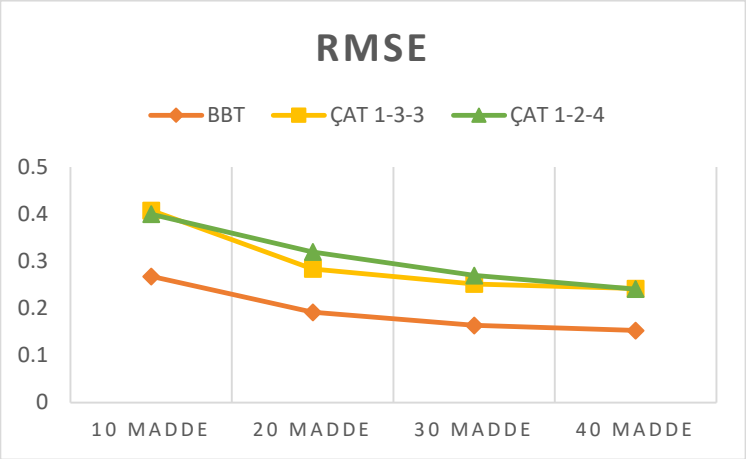
** T simgesi Tukey testinin, D simgesi Dunnet C testinin kullanıldığını göstermektedir.

Sonuç olarak, RMSE ve yanlılık için en düşük değerler tüm test uzunluklarında BBT deseninde elde edilmiş ve bu değerler ÇAT desenlerine ilişkin değerlerden manidar şekilde farklılaşmıştır (Tablo 8). En yüksek RMSE ve yanlılık değerleri ise tüm test uzunluklarında ÇAT 1-2-4 deseninde gözlenmiştir. Grafiklere bakıldığında, ÇAT desenlerine ait değerlerin birbirine yakın olduğu görülmüştür. Ancak Post-Hoc sonuçları incelendiğinde, ÇAT desenlerine ait değerlerin manidar düzeyde farklılaştığı sonucuna ulaşılmıştır. Dolayısıyla, ÇAT 1-2-4 deseninin en yüksek RMSE ve yanlılık değerlerine sahip olduğu ve bu yönden ÇAT 1-3-3 deseninden manidar düzeyde farklılaştığı söylenebilir. Korelasyon içinse tüm test uzunlukları boyunca en yüksek değerler BBT'de, en düşük değerler ise ÇAT 1-2-4 deseninde elde edilmiştir. Desenler arası korelasyon değerleri farkı tüm test uzunlukları boyunca manidar düzeydedir. Tüm sonuçlar birlikte ele alındığında, en düşük RMSE ve yanlılık, en yüksek korelasyon değerlerine sahip olan BBT deseninin en yüksek ölçüm hassasiyetini sağladığı sonucuna ulaşılmıştır. Öte yandan, en yüksek RMSE ve yanlılık, en düşük korelasyon değerlerine sahip olan ÇAT 1-2-4 deseni en düşük ölçüm hassasiyetine sahip desendir.

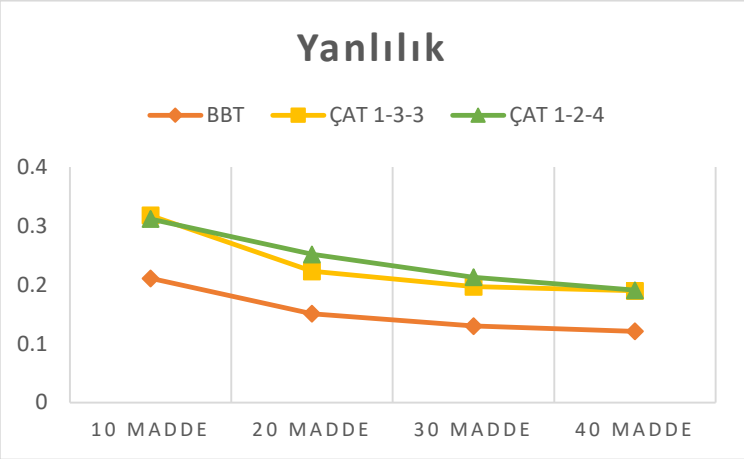
Alt Problem 2'ye İlişkin Bulgu ve Yorumlar

"Testte bulunan DMF'li madde oranı %20 iken, farklı BBT uygulamalarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE ve yanlılık değerleri; test uzunluğu 10, 20, 30 ve 40 olduğu durumda nasıl değişmektedir?"

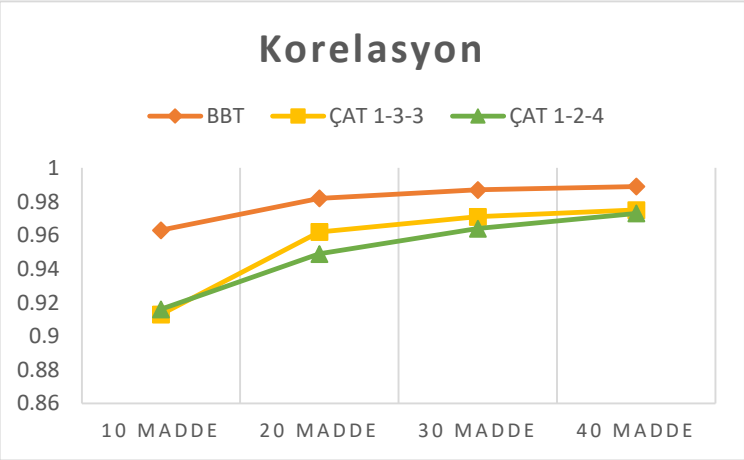
İkinci alt problemde, testte bulunan DMF'li madde oranı %20 iken üç farklı test deseninin etkililiği farklı test uzunlukları altında incelenmiştir. Test desenlerinin etkililiğinin incelenmesinde 30 replikasyondan elde edilen RMSE, yanlılık ve korelasyon değerleri ortalamalarından yararlanılmış ve elde edilen verilerin grafiksel gösterimleri Şekil 12'de sunulmuştur. Tüm test uzunlukları boyunca tüm desenler için hesaplanan değerler detaylı olarak EK-C'de verilmiştir.



Şekil 12a. DMF oranı %20 iken RMSE değerleri



Şekil 12b. DMF oranı %20 iken yanlılık değerleri

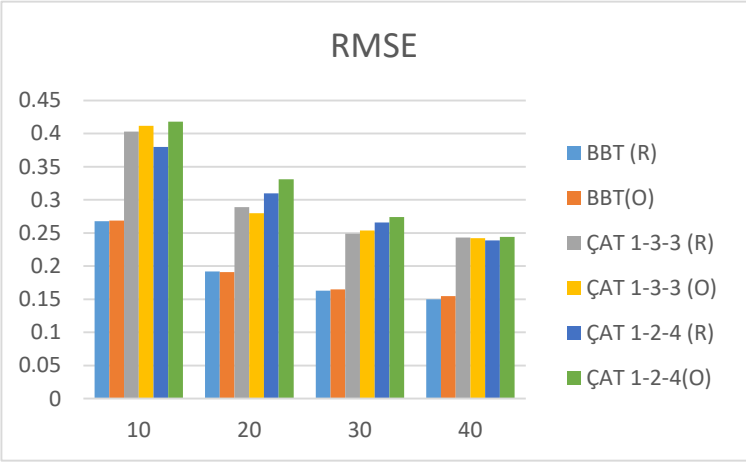


Şekil 12c. DMF oranı %20 iken korelasyon değerleri

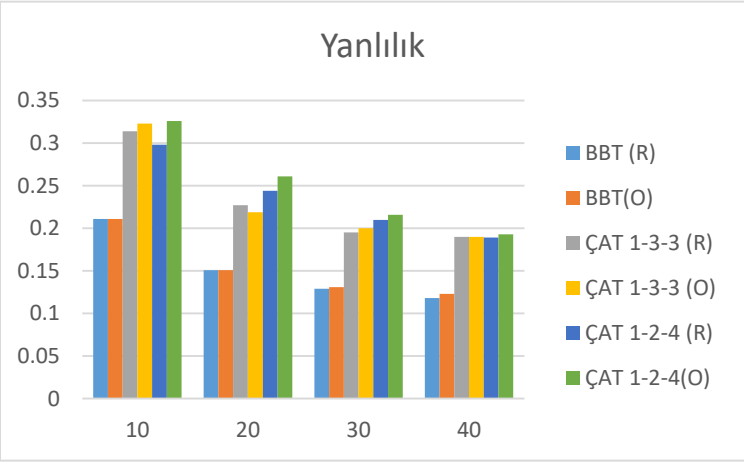
Şekil 12. RMSE, yanlılık ve korelasyon grafikleri (DMF oranı %20)

Şekil 12a'da sunulan RMSE grafiğine göre tüm test uzunlukları için en düşük RMSE değeri BBT'de elde edilmişken; en yüksek RMSE değeri 10 ve 40 maddelik testlerde ÇAT 1-3-3, 20 ve 30 maddelik testler için ÇAT 1-2-4 deseninde ortaya çıkmıştır. Ayrıca, ÇAT 1-3-3 ve 1-2-4 için değerlerin tüm test uzunluklarında birbirine yakın olduğu görülmektedir. BBT'ye ilişkin RMSE değerleri [0.153, 0.268], ÇAT 1-3-3 desenine ait RMSE değerleri [0.242, 0.408] ve ÇAT 1-2-4 desenine ait değerler ise [0.241, 0.400] aralığında değişmiştir (EK-C). Yanlılık değerlerine bakıldığında, BBT deseninin [0.121, 0.211], ÇAT 1-3-3 deseninin [0.190, 0.318] ve ÇAT 1-2-4 deseninin ise [0.191, 0.312] aralığında değerlere sahip olduğu görülmektedir (EK-C). Şekil 11b'de görüldüğü gibi en düşük yanlılık değerleri BBT'ye aittir. En yüksek yanlılık ise 10 madde için ÇAT 1-3-3, diğer test uzunlukları için ÇAT 1-2-4 desenindedir. Test desenlerinin korelasyon açısından karşılaştırıldığı durumda da tüm test uzunlukları boyunca en yüksek korelasyona sahip olan desen BBT iken [0.963, 0.989] , en düşük korelasyon 10 maddelik test için ÇAT 1-3-3, diğer test uzunlukları içinse ÇAT 1-2-4 deseninde elde edilmiştir (EK-C). Tüm desenler için madde sayısı arttıkça korelasyon değerleri de artmış ve birbirine yaklaşmıştır.

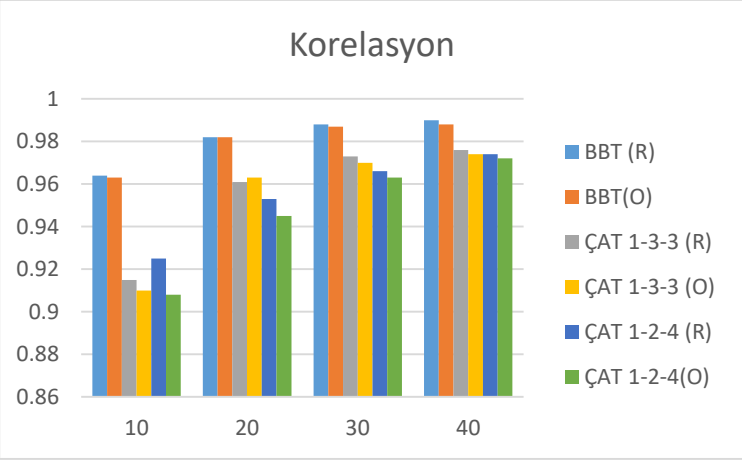
Şekil 12'deki grafiklerden ayrı olarak; odak ve referans gruplarına ilişkin RMSE, yanlılık ve korelasyon değişimlerini detaylı şekilde görebilmek için Şekil 13'te sunulan grafikler çizilmiştir. RMSE ve yanlılık grafikleri incelendiğinde (Şekil 13a-13b), BBT için odak ve referans grupların birbirine oldukça yakın RMSE ve yanlılık değerleri verdiği görülmüştür. ÇAT 1-3-3 için gruplar arası RMSE farkının 10 ve 20 maddelik testte daha yüksek olduğu ancak madde sayısı artışıyla birlikte farkın azaldığı gözlenmiştir. Özellikle 40 maddelik testte değerler birbirine oldukça yakındır. Benzer bir durum yanlılık grafiğinde de gözlenmiştir. ÇAT 1-2-4 deseninde ise, tüm test uzunluklarında odak grup RMSE ve yanlılık değerleri referans grubun değerinden daha yüksektir. Hem RMSE hem de yanlılık için, madde sayısı arttıkça grup değerleri birbirine yaklaşmıştır. Son olarak Şekil 13c'de verilen korelasyon grafiği incelenmiş, BBT deseni için tüm test uzunluklarında referens gruba ait korelasyon değerinin odak gruptakinden daha yüksek olduğu görülmüştür. ÇAT 1-3-3 için de 20 maddelik testte odak grup, diğer test uzunluklarında referans grup daha yüksek korelasyon değerine sahiptir. ÇAT 1-2-4 deseni için yapılan incelemede ise test uzunluğu artıkça gruplar arası farkın azaldığı, ayrıca referans grubun tüm uzunluklarda daha yüksek değere sahip olduğu gözlenmiştir.



Şekil 13a. Gruplara göre RMSE değerleri



Şekil 13b. Gruplara göre yanlılık değerleri



Şekil 13c. Gruplara göre korelasyon değerleri

Şekil 13. Odak ve referans gruplarına göre RMSE, yanlılık ve korelasyon değerleri (DMF oranı %20)

RMSE, yanlılık ve korelasyon değerlerinin test uzunlukları ve test desenleri boyunca manidar düzeyde farklılaşıp farklılaşmadığını incelemek amacıyla tek yönlü ANOVA analizi yapılmıştır. RMSE, yanlılık ve korelasyonun bağımlı değişkenler olarak, test deseninin ise bağımsız değişken olarak alındığı üç ayrı ANOVA analizi yapılmış ve analiz sonuçları her bir test uzunluğu özelinde ayrı ayrı incelenmiştir. Daha önce açıklandığı gibi, korelasyon değerleri Fisher z dönüşümü ile normal dağılıma yaklaştırılmış ve sonrasında analize dahil edilmiştir. RMSE ve yanlılık değerleri için böyle bir dönüşüm yapılmamıştır. ANOVA sonuçları Tablo 9'da sunulmuştur.

Tablo 9

RMSE, Yanlılık ve Korelasyon için Tek Yönlü ANOVA Sonuçları (DMF oranı %20)

		Test Uzunluğu			
Test Deseni		10	20	30	40
RMSE	Test İstatistiği	10154.842 (W)	27820.700 (W)	44714.258 (W)	37739.436 (W)
	p değeri	.000	.000	.000	.000
Yanlılık	Test İstatistiği	21611.658 (W)	25373.859 (W)	31244.173 (W)	18284.355 (W)
	p değeri	.000	.000	.000	.000
Korelasyon	Test İstatistiği	17298.430 (W)	29045.593 (W)	42085.404 (W)	25793.096 (F)
	p değeri	.000	.000	.000	.000

* (F) simgesi varyansların homojenliğinin sağlandığını ve ANOVA F istatistiğinin yorumlandığını
(W) simgesi varyansların homojenliğinin sağlanmadığını ve Welch istatistiğinin yorumlandığını göstermektedir.

Tablo 9'da görüldüğü üzere, RMSE değerlerinin bağımlı değişken olarak alındığı tek yönlü ANOVA analizinde, varyansların homojenliği varsayımı sadece iki durumda sağlanmış, bu durumlarda F diğer durumlarda Welch istatistiği kullanılmıştır. Elde edilen değerin tüm test uzunlukları boyunca manidar düzeyde olması ($p<.05$), RMSE'nin her bir test uzunluğu için test deseni grupları arasında manidar düzeyde farklılık bulunduğunun göstergesidir. Yanlılık ve korelasyonun bağımlı değişken olarak alındığı analiz sonuçları da benzer şekilde sonuçlanmıştır. Hem yanlılık hem de korelasyon değerleri, tüm test uzunluğu alt gruplarında test desenleri arasında manidar düzeyde farklılık göstermektedir. Farklılıkların hangi gruplardan kaynaklandığını tespit etmek için yapılan Post-Hoc karşılaştırması, homojenliğin sağlandığı gruplarda Tukey, sağlanmadığı gruplarda Dunnett C testi ile yapılmış, alınan sonuçlar Tablo 10'da sunulmuştur.

Tablo 10

RMSE, Yanlılık ve Korelasyon için Post-Hoc Sonuçları (DMF oranı %20)

		Ortalama Farkı (I-J)											
		RMSE				Yanlılık				Korelasyon (z)			
Test Deseni (I)	Test Deseni (J)	10 (D)	20 (D)	30 (D)	40 (D)	10 (D)	20 (D)	30 (D)	40 (D)	10 (D)	20 (D)	30 (D)	40 (T)
BBT	ÇAT 1-3-3	-.139*	-.092*	-.088*	-.090*	-.107*	-.072*	-.068*	-.064*	.445*	.371*	.405*	.426*
	ÇAT 1-2-4	-.131*	-.129*	-.106*	-.089*	-.101*	-.101*	-.084*	-.070*	.422*	.522*	.518*	.463*
ÇAT 1-3-3	BBT	.139*	.092*	.088*	.090*	.107*	.072*	.068*	.064*	-.445*	-.371*	-.405*	-.426*
	ÇAT 1-2-4	.008*	-.036*	-.019*	.001	.006*	-.029*	-.016*	-.006	-.023*	.151*	.113*	.037*
ÇAT 1-2-4	BBT	.131*	.129*	.106*	.089*	.101*	.101*	.084*	.070*	-.422*	-.522*	-.518*	-.463*
	ÇAT 1-3-3	-.008*	.036*	.0186*	-.001	-.006*	.029*	.016*	.006	.023*	-.151*	-.113*	-.037*

*. Ortalama Farkı .05 düzeyinde anlamlıdır

** T simgesi Tukey testinin, D simgesi Dunnet C testinin kullanıldığını göstermektedir.

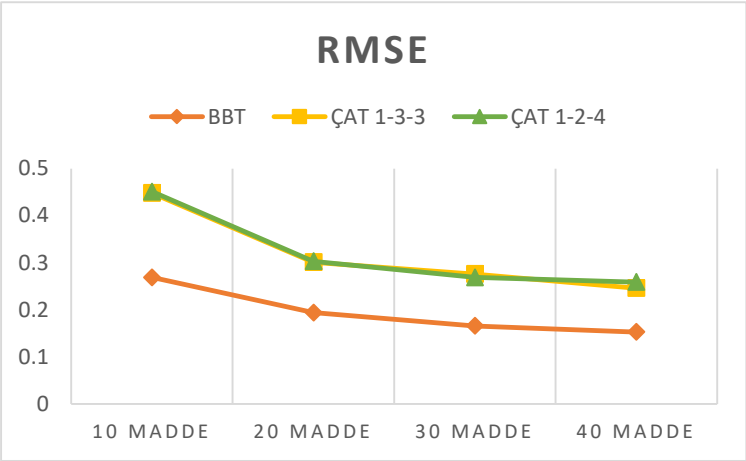
Tablo 10'da verilen Post-Hoc sonuçlarına göre, test desenleri arası RMSE farkı 40 maddelik test için manidar düzeye ulaşmamışken, diğer tüm test uzunluklarında, tüm desenler arasında manidardır. 40 maddelik testte, BBT'ye ait RMSE değeri diğer iki ÇAT deseninden manidar düzeyde farklı olmakla birlikte; ÇAT 1-2-4 ve 1-3-3 desenleri arasında bu fark manidar düzeyde değildir. Desenlere ilişkin yanlılık değerlerinin ortalama farkları incelendiğinde ise, tüm test uzunlukları için üç desenin de birbirinden manidar düzeyde farklılaştığı sonucuna ulaşılmıştır. Yanlılık değerlerinde olduğu gibi korelasyon değerleri için de, tüm test uzunluklarında tüm desen grupları arasında manidar düzeyde fark olduğu görülmüştür.

Sonuç olarak, tüm test uzunlukları boyunca RMSE'nin ve yanlılığın en düşük, korelasyonun ise en yüksek olduğu desen BBT olmuştur. Ayrıca, diğer desenlerle olan RMSE, yanlılık ve korelasyon farkları incelendiğinde (Tablo 10), tüm durumlar için farkın manidar düzeyde olduğu görülmüştür. Dolayısıyla, üç desen arasında en yüksek ölçüm hassasiyetini sağlayan testin BBT olduğu sonucuna varılmıştır. RMSE'nin en yüksek olduğu desen ise 10 ve 40 maddelik testler için ÇAT 1-3-3, 20 ve 30 maddelik testler içinse ÇAT 1-2-4 desenidir. ÇAT desenleri arasındaki fark 10, 20 ve 30 maddelik testlerde manidar düzeye ulaşmıştır. Yanlılığın en yüksek değerleri ise 10 madde için ÇAT 1-3-3, diğer test uzunlukları için ÇAT 1-2-4 desenindedir. En düşük korelasyon 10 maddelik testte ÇAT 1-3-3, diğer test uzunluklarında ise ÇAT 1-2-4 deseninden elde edilmiştir. Tüm sonuçlar birlikte ele alındığında, en düşük ölçüm hassasiyetinin 10 maddelik testte ÇAT 1-3-3, diğer test uzunluklarında ÇAT 1-2-4 deseninde elde edildiği yorumu yapılabilir.

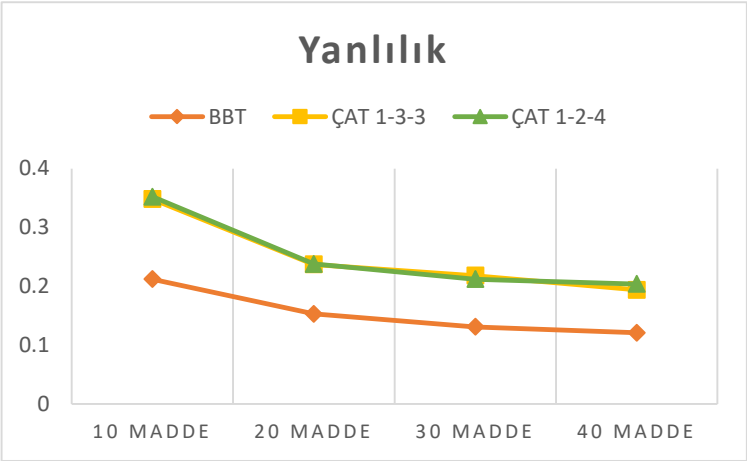
Alt Problem 3'e İlişkin Bulgu ve Yorumlar

"Madde havuzundaki DMF'li madde oranı %30 iken, farklı BBT uygulamalarına (BBT, 1-3-3 ÇAT, 1-2-4 ÇAT) ilişkin RMSE ve yanlılık değerleri; test uzunluğu 10, 20, 30 ve 40 olduğu durumda nasıl değişmektedir?"

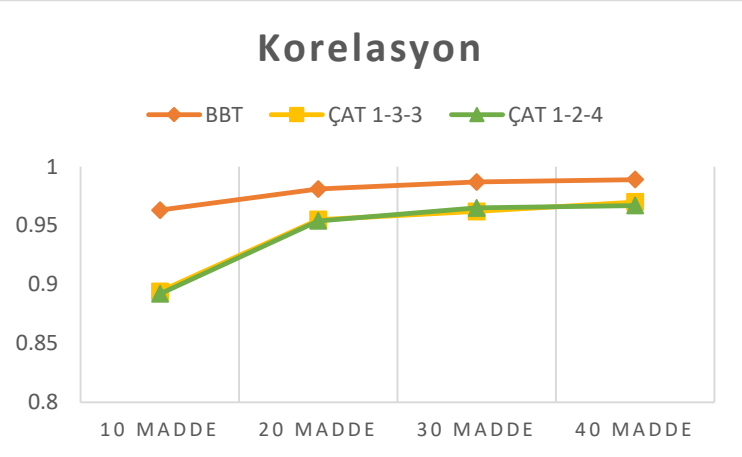
Üçüncü alt problemde, testte bulunan DMF'li madde oranı %30'a çıkarıldığında üç farklı test deseninin performansı farklı test uzunlukları altında incelenmiştir. Test desenlerinin performanslarının değerlendirilmesinde 30 replikasyondan elde edilen RMSE, yanlılık ve korelasyon değerleri ortalamalarından yararlanılmış; elde edilen değerler EK-Ç'de detaylı olarak sunulmuştur. Bu bölümde, bu değerlere ilişkin grafiksel gösterimler sunulmuş (Şekil 14) ve yorumlanmıştır.



Şekil 14a. DMF oranı %30 iken RMSE değerleri



Şekil 14b. DMF oranı %30 iken yanlılık değerleri

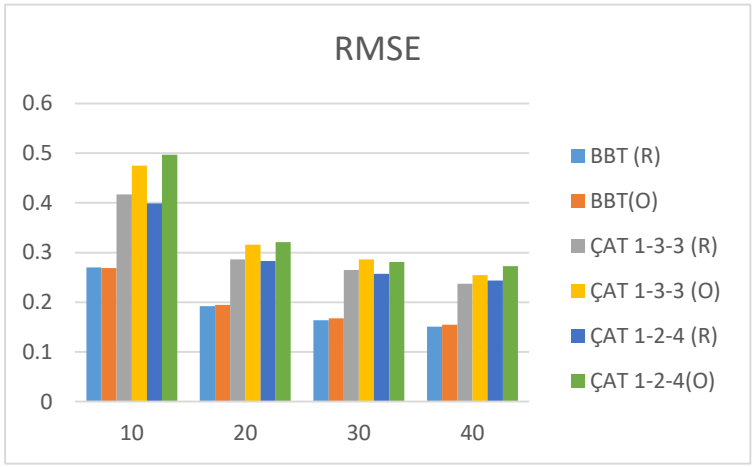


Şekil 14c. DMF oranı %30 iken korelasyon değerleri

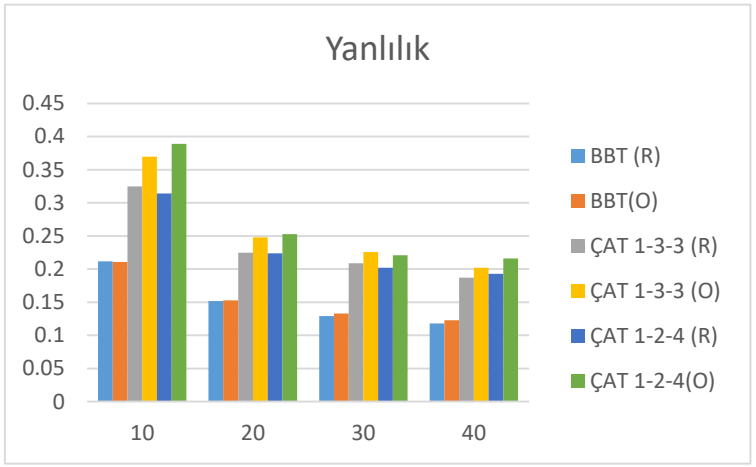
Şekil 14. RMSE, yanlılık ve korelasyon grafikleri (DMF oranı %30)

BBT desenine ilişkin RMSE değerleri [0.153 - 0.269], ÇAT 1-3-3 desenine ait RMSE değerleri [0.246 - 0.448] ve ÇAT 1-2-4 desenine ait değerler ise [0.259 - 0.451] aralığında değişmektedir (EK-Ç). Şekil 14'te de görüldüğü gibi, tüm test uzunlukları için en düşük RMSE değeri BBT deseninde elde edilmiştir. ÇAT 1-3-3 ve 1-2-4 desenlerine ilişkin RMSE değerleri tüm test uzunlukları için birbirine oldukça yakın görülmektedir (Şekil 14a). Yanlılık değerleri incelendiğinde, BBT deseninin [0.121-0.212], ÇAT 1-3-3 deseninin [0.194-0.348] ve ÇAT 1-2-4 deseninin ise [0.204-0.352] aralığında değerlere sahip olduğu gözlenmiştir (EK-Ç). Şekil 14b'de görüldüğü gibi en düşük yanlılık değerleri tüm test uzunluklarında BBT deseninde hesaplanmıştır. ÇAT desenleri ise birbirine oldukça yakın sonuçlar vermiştir. Test uzunluğu arttıkça tüm desenler için yanlılık değeri azalmış ve desenler arası yanlılık değerleri farkı azalmıştır. Test desenleri korelasyon değerleri açısından karşılaştırıldığında ise, tüm test uzunlukları boyunca en yüksek korelasyon değerine sahip olan desen BBT desenidir [0.963-0.989] (Şekil 14c). ÇAT desenlerine ait korelasyon değerleri tüm test uzunluklarında birbirine oldukça yakın görülmektedir (Şekil 14c). Tüm desenler için madde sayısı arttıkça korelasyon değerleri de artmıştır.

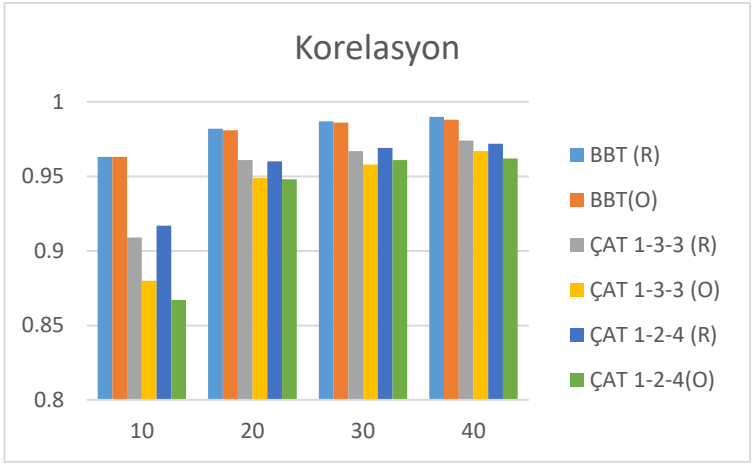
Şekil 14'te verilen genel grup grafiklerine ek olarak daha detaylı yorum yapabilmek için odak ve referans gruplarına ilişkin RMSE, yanlılık ve korelasyon grafikleri de Şekil 15'te sunulmuştur. Şekil 15a'da sunulan RMSE grafiğine bakıldığında, BBT için tüm test uzunluklarında küçük farklar olduğu ve odak grubun daha yüksek değere sahip olduğu gözlenmiştir. ÇAT 1-3-3 ve 1-2-4 desenlerinin her ikisinde de test uzunluğu arttıkça gruplar arası fark azalmış ancak tüm test uzunluklarında odak grup daha yüksek RMSE değeri almıştır. Ayrıca ÇAT desenlerindeki gruplar arası fark, BBT desenindeki farktan tüm test uzunluklarında daha yüksektir. Yanlılık değerleri için de aynı yorumlar yapılabilir (Şekil 15b). Son olarak Şekil 15c'de verilen korelasyon grafiği incelenmiş, özellikle ÇAT desenlerinde 10 maddelik testte gruplar arası farkın büyük olduğu gözlenmiştir. Tüm desenlerde ve tüm test uzunluklarında referans grubun korelasyon deseni odak grubun değerinden daha yüksektir.



Şekil 15a. Gruplara göre RMSE değerleri



Şekil 15b. Gruplara göre yanlılık değerleri



Şekil 15c. Gruplara göre korelasyon değerleri

Şekil 15. Odak ve referans gruplarına göre RMSE, yanlılık ve korelasyon değerleri (DMF oranı %30)

RMSE, yanlılık ve korelasyon değerlerinin test uzunlukları ve test desenleri boyunca manidar düzeyde farklılaşıp farklılaşmadığını incelemek amacıyla tek yönlü ANOVA analizi yapılmıştır. RMSE, yanlılık ve korelasyonun bağımlı değişkenler olarak, test deseninin ise bağımsız değişken olarak alındığı üç ayrı ANOVA analizi yapılmış ve analiz sonuçları her bir test uzunluğu özelinde ayrı ayrı incelenmiştir. Daha önceki ANOVA analizlerinde olduğu gibi korelasyon değerleri doğrudan kullanılmamış, Fisher z dönüşümüyle elde edilen değerler analize alınmıştır. Elde edilen analiz bulguları Tablo 11'de sunulmuştur.

Tablo 11

RMSE, Yanlılık ve Korelasyon için Tek Yönlü ANOVA Sonuçları (DMF oranı %30)

		Test Uzunluğu			
Test Deseni		10	20	30	40
RMSE	Test İstatistiği	27142.360 (W)	19282.802 (W)	40663.748 (W)	35739.601 (W)
	p değeri	.000	.000	.000	.000
Yanlılık	Test İstatistiği	26213.056 (W)	19201.607 (W)	37629.012 (W)	42555.788 (W)
	p değeri	.000	.000	.000	.000
Korelasyon	Test İstatistiği	11845.330 (W)	15829.182 (F)	25805.746 (W)	21931.792 (F)
	p değeri	.000	.000	.000	.000

* (F) simgesi varyansların homojenliğinin sağlandığını ve ANOVA F istatistiğinin yorumlandığını
(W) simgesi varyansların homojenliğinin sağlanmadığını ve Welch istatistiğinin yorumlandığını göstermektedir.

Tablo 11'e bakıldığında; RMSE, yanlılık ve korelasyon için hiçbir test uzunluğu düzeyinde varyansların homojenliği varsayımının sağlanmadığı ve bu nedenle Welch istatistiğinin yorumlandığı görülmektedir. Her bir bağımlı değişken için tüm test uzunlukları boyunca elde edilen anlamlılık değerlerinin .05'ten küçük olması, gruplar arası farklılıkların manidar düzeyde olduğunu göstermiştir. Yani; RMSE, yanlılık ve korelasyon değerleri her bir test uzunluğu için tüm desenler arasında manidar düzeyde farklılık göstermiştir. Farklılıkların hangi gruplardan kaynaklandığını tespit etmek için yapılan Post-Hoc karşılaştırması, homojenliğin sağlandığı gruplarda Tukey, sağlanmadığı gruplarda Dunnett C testi ile yapılmış, alınan sonuçlar Tablo 12'de sunulmuştur.

Tablo 12

RMSE, Yanlılık ve Korelasyon için Post-Hoc Sonuçları (DMF oranı %30)

Test Deseni (I)	Test Deseni (J)	Ortalama Farkı (I-J)											
		RMSE				Yanlılık				Korelasyon (z)			
		10 (D)	20 (D)	30 (D)	40 (D)	10 (D)	20 (D)	30 (D)	40 (D)	10 (D)	20 (T)	30 (D)	40 (T)
BBT	ÇAT 1-3-3	-.178*	-.107*	-.110*	-.093*	-.136*	-.083*	-.086*	-.074*	.368*	.365*	.361*	.362*
BBT	ÇAT 1-2-4	-.182*	-.109*	-.103*	-.106*	-.140*	-.085*	-.080*	-.084*	.387*	.479*	.444*	.425*
ÇAT 1-3-3	BBT	.178*	.107*	.110*	.093*	.136*	.083*	.086*	.074*	-.368*	-.365*	-.361*	-.362*
ÇAT 1-3-3	ÇAT 1-2-4	-.003*	-.001	.007*	-.013*	-.004*	-.002*	.006*	-.010*	.019*	.114*	.084*	.063*
ÇAT 1-2-4	BBT	.182*	.109*	.103*	.106*	.140*	.085*	.080*	.084*	-.387*	-.479*	-.444*	-.425*
ÇAT 1-2-4	ÇAT 1-3-3	.003*	.001	-.007*	.013*	.004*	.002*	-.006*	.010*	-.019*	-.114*	-.084*	-.063*

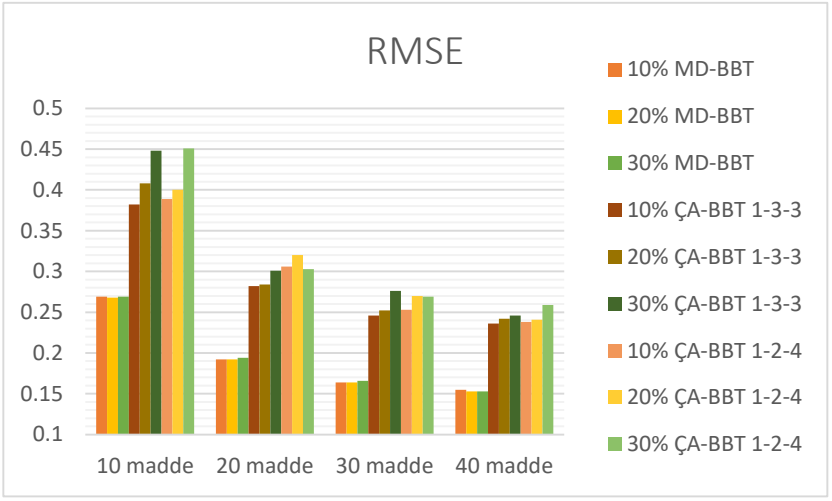
*. Ortalama Farkı .05 düzeyinde anlamlıdır

** T simgesi Tukey testinin, D simgesi Dunnet C testinin kullanıldığını göstermektedir.

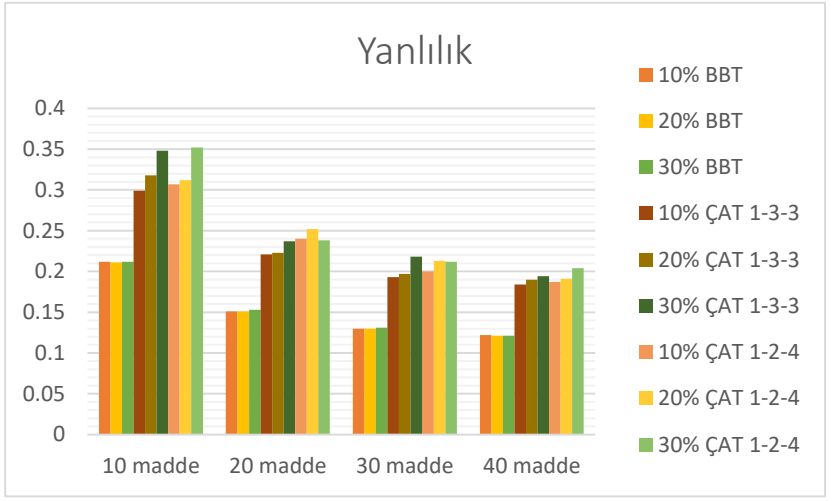
Tablo 12'de verilen Post-Hoc karşılaştırması sonuçları RMSE değişkeni açısından incelendiğinde, 10, 30 ve 40 maddelik testler için tüm desenlerin arasında manidar düzeyde fark olduğu; ancak 20 maddelik testte ÇAT 1-3-3 ve 1-2-4 desenleri arasındaki .001'lik ortalama farkın manidar düzeye ulaşmadığı görülmektedir. Desenlere ilişkin yanlılık değerlerinin ortalama farklarına bakıldığında ise, tüm test uzunlukları için üç desenin de birbirinden manidar düzeyde farklılaştığı sonucuna ulaşılmıştır. Son olarak korelasyon değerleri Post-Hoc sonuçları incelenmiş ve tüm test uzunlukları için tüm desenlerin arasında manidar düzeyde fark olduğu sonucuna ulaşılmıştır.

Sonuç olarak, diğer alt problemlerde olduğu gibi, BBT deseni tüm test uzunlukları boyunca en düşük RMSE ve yanlılık, en yüksek korelasyon değerlerinin görüldüğü desen olmuştur. Ayrıca, tüm durumlarda diğer desenlerden manidar düzeyde farklılık göstermiştir (Tablo 12). Dolayısıyla en yüksek ölçüm hassasiyetine sahip desen BBT deseni olmuştur ve bu hassasiyet tüm test uzunlukları boyunca diğer desenlerin sahip olduğu hassasiyetten manidar düzeyde yüksektir. Grafiklere bakıldığında, ÇAT desenlerine ait değerlerin birbirine oldukça yakın olduğu görülmüştür. Ancak detaylı incelendiğinde, 30 madde haricindeki diğer test uzunluklarında en yüksek RMSE ve yanlılık, en düşük korelasyon değerleri ÇAT 1-2-4 deseninde elde edilmiştir. Buna göre, 30 maddelik test için en düşük ölçüm hassasiyeti ÇAT 1-3-3 deseninde, diğer durumlar için ÇAT 1-2-4 deseninde gözlenmiştir.

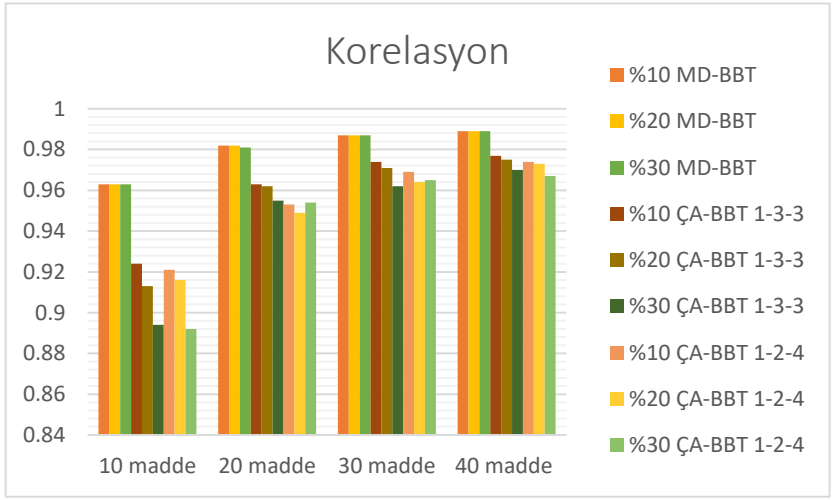
Son olarak, testte bulunan DMF'li madde oranındaki artışın RMSE, yanlılık ve korelasyon üzerindeki etkilerini betimsel olarak görmek amacıyla grafikler oluşturulmuş ve Şekil 16'da sunulmuştur.



Şekil 16a. Gruplara göre RMSE değerleri



Şekil 16b. Gruplara göre yanlılık değerleri



Şekil 16c. Gruplara göre korelasyon değerleri

Şekil 16. DMF oranlarına göre RMSE, yanlılık ve korelasyon grafikleri

RMSE grafiğine bakıldığında (Şekil 16a), BBT için RMSE değerlerinin farklı DMF oranlarında birbirine oldukça yakın olduğu; ancak ÇAT desenlerinde, özellikle test uzunluğunun en düşük olduğu 10 maddelik testte, DMF oranı artışının RMSE değerini artırdığı görülmüştür. ÇAT desenleri için Madde sayısı arttıkça DMF oranı artışının RMSE üzerindeki etkisi düşmüştür. Benzer şekilde yanlılık değerleri de BBT için farklı DMF oranlarında birbirine yakındır (Şekil 16b). ÇAT desenlerinde ise 10 maddelik testte DMF oranı artışı yanlılık değerlerini oldukça etkilemiş, test uzunluğu artıkça bu etki azalmıştır. Şekil 16c'deki korelasyon grafiğine bakıldığında da RMSE ve yanlılık için yapılan yorumlara benzer yorumlar yapılabilir. ÇAT desenleri için DMF oranı arttıkça korelasyon değerlerinin azaldığı ve bu azalmanın en ciddi etkiyi 10 maddelik testte gösterdiği belirlenmiştir. BBT içinse DMF oranı artışı çok büyük bir etki oluşturmamıştır.

Bu bulgular, BBT performansının testteki DMF'li madde oranı artışından en az etkilenen test deseni olduğunu göstermiştir. ÇAT desenleri genellikle birbirine paralel durumlar sergilemiştir. Bu desenler özellikle 10 maddelik testte DMF oranı değişiminden etkilenmiş, test uzunluğu arttıkça bu etki azalmıştır.

Bölüm 5
Sonuç, Tartışma ve Öneriler

Bu çalışma kapsamında, testte değişen madde fonksiyonu (DMF) gösteren maddelerin yer almasının, bireyselleştirilmiş bilgisayarlı (BBT) ve çok aşamalı testlerin (ÇAT) etkililiği üzerindeki etkisinin farklı koşullar altında incelenmesi amaçlanmıştır. Bu amaçla, simülasyon yöntemiyle veriler üretilmiş ve farklı test desenlerinin performansları (BBT, ÇAT 1-3-3 ve ÇAT 1-2-4) farklı test uzunlukları ve DMF oranları altında karşılaştırılmıştır. Elde edilen bulgular bir önceki bölümde detaylı bir şekilde açıklanmıştır. Bu bölümde, araştırma sonucunda elde edilen bulgular özetlenmiş ve sonrasında literatürdeki benzer çalışmalarla birlikte tartışılmıştır. Son olarak ileriki araştırmalara yönelik önerilere yer verilmiştir.

Sonuçlar

Araştırma kapsamında farklı test desenlerinin (BBT, 1-3-3 ÇAT ve 1-2-4 ÇAT) performansları farklı koşullar altında (testteki DMF'li madde oranı, test uzunluğu) incelenmiştir. Test performanslarının değerlendirilmesinde RMSE, yanlılık ve korelasyon değerleri birlikte ele alınmıştır. Elde edilen sonuçlar RMSE açısından incelendiğinde, tüm koşullar için BBT deseninin en düşük RMSE değerine sahip olduğu görülmüştür. Ayrıca ANOVA analizi sonuçları, BBT'ye ilişkin RMSE değerlerinin diğer desenlerinkinden manidar düzeyde farklılaştığını göstermiştir. ÇAT 1-3-3 ve 1-2-4 desenleri karşılaştırıldığında ise, genel bir yorum yapılamamaktadır. DMF oranının %10 olduğu koşulda, tüm test uzunlukları boyunca 1-2-4 desenine ait RMSE değeri 1-3-3 deseninden manidar düzeyde yüksek iken; %20 olduğu koşulda 10 ve 40 maddelik testte, %30 olduğu durumda 30 maddelik testte 1-3-3 deseni daha yüksek RMSE değerleri vermiştir. Ayrıca tüm desen ve tüm koşullarda, madde sayısının artışı ile birlikte RMSE değerinde azalma meydana gelmiştir. Yanlılık değerleri incelendiğinde, BBT'nin her koşulda en düşük değere sahip olduğu görülmüştür. RMSE'ye benzer olarak yanlılık için de DMF oranının %10 olduğu koşulda, tüm test uzunlukları boyunca 1-2-4 desenine ait yanlılık değeri 1-3-3 deseninden manidar düzeyde yüksek iken; %20 olduğu koşulda 10 ve 40 maddelik testte, %30 olduğu durumda 30 maddelik testte 1-3-3 deseni daha yüksek yanlılık değerleri vermiştir. Ayrıca tüm desen ve tüm koşullarda, madde sayısının artışı ile birlikte yanlılık değerinde azalma meydana gelmiştir. Son olarak korelasyon

değerlerine bakılarak edinilen bulgular BBT'nin her koşulda en yüksek korelasyon değerine sahip olduğunu göstermiştir. En düşük korelasyon değerleri ise; %20 DMF - 10 madde, %30 DMF – 30 madde koşullarında 1-3-3 deseninde diğer tüm koşullarda ise 1-2-4 deseninde elde edilmiştir. Ayrıca korelasyon değerlerinin tüm koşullarda test uzunluğu arttıkça arttığı görülmüştür. DMF oranı artışının test desenlerinin performanslarını ne şekilde etkilediğine bakıldığında ise, BBT'nin DMF'li madde oranına bağlı olmaksızın RMSE, yanlılık ve korelasyon yönünden benzer sonuçlar verdiği görülmüştür (Şekil 16). Ancak ÇAT desenleri için aynı durum söz konusu değildir. ÇAT desenlerinde DMF oranının artışı genellikle RMSE ve yanlılık değerlerinde artışa ve korelasyon değerlerinde düşüşe yol açmıştır. Özellikle de 10 maddelik testler, DMF oranının artışından BBT'ye göre daha fazla etkilenmiş, madde sayısı arttıkça bu etki azalmıştır.

Yukarıdaki bilgiler birlikte yorumlandığında; tüm test uzunluğu ve DMF oranı koşullarında BBT'nin diğer iki ÇAT desenine kıyasla daha iyi ölçüm hassasiyeti sağladığı yorumu yapılabilir. Ayrıca DMF'li madde oranı artışından en az etkilenen desen BBT'dir. Dolayısıyla, diğer desenlere kıyasla BBT DMF'nin etkisini daha fazla azaltabilmiştir. İki ÇAT deseni kendi arasında kıyaslandığında, 1-3-3 deseninin daha fazla sayıda koşulda yüksek ölçüm hassasiyeti sunduğu görülmüştür. Ancak eldeki bulgular 1-3-3 deseninin 1-2-4 desenine kıyasla daha yüksek performans gösterdiğini söylemek için yeterli değildir.

Tartışma

Bu çalışmadan elde edilen temel bulgu, tüm test uzunlukları boyunca DMF etkisini en aza indirgeyen desenin BBT deseni olduğudur. BBT'nin DMF'nin etkisini düzenleyebildiği bulgusu Piromsombat'ın (2014) çalışmasından elde edilen bulgularla paralellik göstermektedir. Piromsombat testte bulunan DMF'li maddelerin yetenek kestirimi üzerindeki etkisini BBT'ler üzerinde incelemiş; DMF'li maddelerin testin başlarında gelmesi durumunda, özellikle de DMF düzeyi orta düzeyde olduğunda, BBT'nin DMF etkisini düzenleyebildiğini ortaya koymuştur. Diğer durumlarda da, BBT DMF'nin etkisini azaltmıştır. Çalışmadan elde edilen bir diğer bulgu, DMF oranı artışının BBT performansı üzerindeki etkisinin, ÇAT desenlerindeki etkisine kıyasla daha düşük olduğudur. ÇAT desenleri özellikle madde sayısının 10 olduğu durumda DMF oranı artışından oldukça etkilenmiştir.

BBT'lerde uyarlama noktası sayısının ÇAT'lara göre daha fazla olması BBT ölçüm hassasiyetinin daha yüksek olmasını sağlayabilmektedir (Sarı, 2016; Tay, 2015). Örneğin, 1-3-3 panel deseninde madde sayısı fark etmeksizin yalnızca iki uyarlama noktası bulunurken 20 maddelik bir BBT'de 19 uyarlama noktası bulunmaktadır. Dolayısıyla, testte DMF'li madde bulunması durumunda BBT'nin ÇAT desenlerine kıyasla daha iyi ölçüm hassasiyeti sunması ve DMF oranından daha az etkilenmesi beklenen bir sonuçtur. Literatürde, DMF'li maddelerin BBT'ler üzerindeki etkisini inceleyen başka çalışmalara rastlanmadığı için bu bulguya ilişkin tartışma sınırlı kalmıştır.

DMF etkisinden bağımsız olarak BBT ve ÇAT desenlerinin karşılaştırıldığı çalışmalar da incelenmiştir. Kim ve Plake (1993), BBT'nin ve ÇAT'ın ölçüm hassasiyetini ilk aşama modül uzunluğu (10, 15 ve 20 madde), toplam test uzunluğu (40, 45 ve 50 madde), ikinci aşama modül sayısı (6, 7, 8 modül) ve ilk aşama modülündeki madde güçlüğü dağılımı koşullarında incelemiş; BBT'nin ÇAT'a göre ölçüm hassasiyeti noktasında daha iyi sonuç verdiğini ortaya koymuştur. Patsula (1999) tarafından yapılan çalışmada; farklı BBT, kağıt-kalem testi ve ÇAT desenlerinden (aşama sayısı, her aşamadaki modül sayısı, her modüldeki madde sayısı) elde edilen yetenek kestirimlerinin doğrulukları karşılaştırılmış ve BBT'lerin en doğru yetenek kestirimini ürettiği ve her aşamadaki modül sayısının artmasının ölçüm hassasiyeti ve etkililiğini etkilediği ortaya konmuştur. Bir diğer çalışmada ise Sarı (2016), farklı uzunluktaki testlerde içerik alanları sayısı değişmekte iken BBT ve ÇAT'tan elde edilen sonuçların hassasiyetini araştırmıştır. Çalışmanın ana bulgusu, tüm koşullar için BBT'nin diğer iki ÇAT desenine göre daha iyi sonuç vermiş ve iki ÇAT deseninin ise karşılaştırılabilir sonuçlar sunmuş olduğudur. Ayrıca Tay (2015), BBT'lerin ÇAT'lara nazaran daha fazla uyarlama noktasına sahip olduğunu bu nedenle daha etkili desenler olduğunu belirtmiştir. Literatürdeki çalışmalardan elde edilen ortak sonuç, farklı çalışmalarda ve farklı koşullar altında BBT'nin ÇAT'a göre daha iyi sonuç vermesidir. Bu çalışmalardan yola çıkılarak yapılan bu çıkarım, çalışma sonucunda elde edilen BBT performansının ÇAT'a göre daha yüksek olduğu bulgusuyla paralellik göstermektedir.

Eldeki bulgular incelendiğinde, tüm desenler için test uzunluğu arttıkça RMSE ve yanlılık değerlerinin azaldığı ve korelasyon değerlerinin arttığı görülmüştür. Dolayısıyla test uzunluğunun artmasının ölçüm hassasiyetini artırdığı sonucuna

ulaşılabilir. Bu bulguya benzer şekilde, Sarı (2016) da çalışmasında test uzunluğunu artırmanın hem BBT hem de ÇAT için RMSE ve yanlılık değerinin düşüşü ve korelasyonun artışıyla sonuçlandığını ortaya koymuştur.

Araştırma sonucunda elde edilen bir diğer bulgu; ÇAT desenlerinin kendi arasında karşılaştırılmasına ilişkin 1-3-3 deseninin daha fazla sayıda koşulda yüksek ölçüm hassasiyeti sunduğu ancak eldeki bulguların 1-3-3 deseninin 1-2-4 desenine kıyasla daha yüksek performans gösterdiğini söylemek için yeterli olmadığıdır. Literatürde, bu iki desen arasında karşılaştırma yapılan bir çalışmaya rastlanamamıştır. İlgili bulguya ilişkin bir tartışma yapabilmek için farklı çalışmalardan elde edilen bulgulara ihtiyaç duyulmaktadır.

Öneriler

Uygulayıcılara yönelik öneriler. Araştırma sonucundan elde edilen bulgulara yönelik öneri şu şekildedir;

- Testte DMF'li maddelerin bulunduğu durumlar için BBT'nin ÇAT'a kıyasla daha iyi sonuç verdiği görülmüştür. Bu çalışmayla benzer koşulları taşıyan durumlarda BBT'nin kullanımı önerilebilir.

- ÇAT desenleri DMF'li maddelerden BBT'ye göre daha fazla etkilenmiştir. Kullanılan her iki desen de DMF etkisini düzenleyememiş, ölçüm hassasiyeti BBT'ye kıyasla daha negatif etkilenmiştir. Eğer ÇAT kullanılacaksa DMF analizleri mutlaka yapılmalıdır.

- Özellikle test uzunluğu 10 madde olduğu durumda DMF oranı artışı ÇAT'ların ölçüm hassasiyetini olumsuz etkilemiştir. Bu durumlarda ÇAT kullanımı tercih edilmemeli veya oldukça dikkatli kullanılmalıdır.

- RMSE, yanlılık ve korelasyon grafiklerinde (Şekil 12, 14, ve 16), tüm desenler için özellikle 30 maddeden sonra eğimin azaldığı ve grafik çizgilerinin düzleşmeye başladığı söylenebilir. Dolayısıyla, DMF varlığından şüphelenildiği durumlarda en az 30 maddelik testin kullanılması önerilebilir.

Araştırma yapacaklara yönelik öneriler. İleriki araştırmalara yönelik öneriler aşağıda listelenmiştir;

- Araştırma kapsamında kullanılan veri seti simülasyon verisi ile sınırlıdır. İleriki araştırmalarda gerçek veri seti ile çalışılması önerilebilir.

- Çalışmada kullanılan madde havuzu, araştırmacı tarafından belirlenen madde parametreleri ile sınırlıdır. Farklı madde parametre dağılımları ve değerleriyle madde havuzu oluşturulup çalışma tekrarlanabilir.

- Çalışma kapsamında yalnızca ikili puanlanan maddeler dikkate alınmıştır. Çok kategorili puanlanan maddeler ile benzer çalışmalar yapılabilir.

- Çalışmanın sınırlılıklarından biri; DMF gösteren maddelerin BBT için rastgele dağılırken, ÇAT için ikinci aşamadaki modüle sabitlenmiş olmasıdır. DMF gösteren maddeler farklı sıralamalarda denenebilir.

- DMF'li maddeler üretilirken yalnızca tek biçimli DMF gösterecek şekilde üretilmiştir. Tek biçimli olmayan DMF gösteren maddelerin eklenmesiyle benzer çalışmalar yapılabilir.

- Üretilen DMF'li maddelerin etki büyüklükleri değiştirilerek çalışma tekrarlanabilir.

- Bu çalışmada test sonlandırma kuralı olarak yalnızca sabit uzunluk kullanılmıştır. Farklı test sonlandırma kuralları kullanılarak çalışma tekrarlanabilir.

- Çalışmanın bir diğer sınırlılığı, ÇAT için yalnızca iki desen kullanılmış olmasıdır. Farklı test desenleri ile çalışma tekrarlanabilir.

Kaynaklar

Aksu-Dünya, B. (2017). *Item parameter drift in computer adaptive testing due to lack of content knowledge within sub-populations* (Doctoral Dissertation). University of Illinois, Chicago.

Angoff, W. H. & Huddleston, E. M. (1958). *The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test*. Princeton, N.J.: Educational Testing Service.

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

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565-580. http://dx.doi.org/10.1177/0146621612455090

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

Barton, M. A. & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1981*(1), i-8.

Berger, S., Verschoor, A. J., Eggen, T. J. H. M. & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, *4*(1), 1–18. Retrieved from https://doi.org/10.3389/feduc.2019.00001.

Birdsall, M. (2011). Implementing computer adaptive testing to improve achievement opportunities. *Office of Qualifications and Examinations Regulation Report*, April 2011. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606023/0411_MichaelBirdsall_implementing-computer-testing-_Final_April_2011_With_Copyright.pdf

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: Lord, F.M. and Novick, M.R. (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, Mass.: Addison-Wesley.

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: Item banking and computerized adaptive assessment. *Quality of Life Research*, *16*(1), 95-108.

Boztunç-Öztürk, N. (2019). How the length and characteristics of routing module affect ability estimation in CA-MST? *Universal Journal of Educational Research*, *7*(1), 164–170.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items (4th* ed.). Sage Publications, Inc.

Chang, H.H. & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*(1), 37–52.

Chen, W.H. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Chu, M.W. & Lai, H. (2013). *Detecting biased items using CATSIB to increase fairness in computer adaptive tests*. *Alberta Journal of Educational Research*, *59*(4), 630–643.

Clauser, B. E. & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Corey, D. M., Dunlap, W. P. & Burke, M. J. (1998) Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations. *The Journal of General Psychology*, *125* (3), 245-261, doi: 10.1080/00221309809595548

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont CA: Wadsworth Group/Thomson Learning.

Dai, S., Wang, X. & Svetina, D. (2019). Package 'subscore': Computing subscores in classical test theory and item response theory. Retrieved from https://cran.r-project.org/web/packages/subscore/subscore.pdf

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

Deveci-Ateşok, N. (2008). *Üniversitelerarası kurul yabancı dil sınavının madde yanlılığı bakımından incelenmesi* (Doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Doğruöz, E. (2018). *Bireyselleştirilmiş çok aşamalı testlerin test birleştirme yöntemlerine göre incelenmesi* (Doktora tezi). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah N.J.: L. Erlbaum Associates.

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

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 37–60). Mahwah N.J.: Lawrence Erlbaum Associates.

Fraenkel, J. R., Wallen, N. E. & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). New York: McGraw-Hill.

Gierl, M. J., Lai, H. & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2-3), 188–203. Retrieved from https://doi.org/10.1080/13803611.2013.767622

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

Hambleton, R. K., Jac, N. Z. & Pieters, J. P. M. (2000). Computerized adaptive testing: Theory, applications and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications* (4. ed., pp. 341–366). New York: Springer.

Hambleton, R. K. & Swaminathan, H. (1991*). Item response theory: Principles and applications*. New York: Springer.

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

Han K. T., Guo F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing (GMAC® Research Reports, RR-11-02). Retrieved from https://pdfs.semanticscholar.org/ac5f/cb8d43575b3e1661e51a08c060e353712509.pdf

Harwell, M., Stone, C. A., Hsu, T. C. & Kirisci, L. (2016). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice,* Summer 2007, 44-52.

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

Kim, H. & Plake, B. S. (1993). *Monte carlo simulation comparison of two-stage testing and computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993). Retrieved from https://files.eric.ed.gov/fulltext/ED357041.pdf

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.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359 – 375.

Lamoré, M. (2017). *Comparing a multistage and a linear summative test on ability estimate precision and classification accuracy* (Master Thesis). University of Twente.

Lei, P. W., Chen, S. Y. & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, *43*(3), 245-264. Retrieved from http://dx.doi.org/10.1111/j.1745-3984.2006.00015.x

Lord, F. M. (1968). Some test theory for tailored testing. *ETS Research Bulletin Series*, *1968*(2), i-62. Retrieved from https://doi.org/10.1002/j.2333-8504.1968.tb00562.x

Lord, F. M. (1971). A theoretical study of the measurement effectiveness of flexilevel tests. *ETS Research Report Series*, *1971*(1), i-13. Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1971.tb00185.x

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, New Jersey: Routledge.

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

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

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249.

Luecht, R. M. & Sireci, S. G. (2011). A review of models for computer-based testing. *College Board Research Reports*, *2011*(12). Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Luo, X. (2018). Package 'xxIRT': Item response theory and computer-based testing in R. Retrieved from https://cran.r-project.org/web/packages/xxIRT/xxIRT.pdf

Magis, D., Raiche, G. & Barrada, J. R. (2018). Package 'catR': Generation of IRT response patterns under computerized adaptive testing. Retrieved from https://cran.r-project.org/web/packages/catR/catR.pdf

Magis, D., Yan, D. & von Davier, Alina, A. (2018). Package 'mstR': Procedures to generate patterns under multistage testing. Retrieved from https://cran.r-project.org/web/packages/mstR/mstR.pdf

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

Makransky, G. & Glas, C. A. W (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement*, *46* (2013), 3228-3237.

Morris, T. P., White, I. R. & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

National Research Council (1999). *Designing mathematics or science curriculum programs: A guide for using mathematics and science education standards*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog/9658.html

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive testing. *Journal of the American Statistical Association*, *70*, 351-356.

Öğretmen, T. ve Doğan, N. (2004). OKÖSYS Matematik alt testine ait maddelerin yanlılık analizi. *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, *5*(8), 2–12.

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing* (Doctoral Dissertation). University of Massachusetts, Amherst.

Piromsombat, C. (2014). *Differential item functioning in computerized adaptive testing: Can cat self-adjust enough*? (Doctoral Dissertation). University of Minnesota.

R Core Team. (2018). R: A language and environment for statistical computing: R foundation for statistical computing.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Reise, S. P., Cook, K. F. & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment*. Routledge.

Revelle, W. (2019). Package 'psych': Procedures for Psychological, Psychometric, and Personality Research. Retrieved from https://cran.r-project.org/web/packages/psych/psych.pdf

Rome, L. (2017). *Evaluating item selection methods for adaptive tests with complex content constraints* (Doctoral Dissertation). The University of Wisconsin, Milwaukee.

Rudner, L. M. (2010). Implementing the graduate management admission test computerized adaptive test. In W. J. van der Linden ve C. A.W. Glas (Eds.), *Elements of Adaptive Testing*. Springer.

Sarı, H. İ. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. Computerized multistage testing* (Doctoral Dissertation). University of Florida.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*(1), 146-148.

Steinberg, L., Thissen, D. & Wainer, H. (2000). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed., p. 185–229). Mahwah N.J.: Lawrence Erlbaum Associates.

Tabachnick, B. G., Fidell, L. S. & Ullman, J. B. (2007). *Using multivariate statistics*. Pearson Boston, MA.

Tay, P. H. (2015). *On-the-fly assembled multistage adaptive testing* (Doctoral Dissertation). University of Illinois.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed.). Mahwah N.J.: Lawrence Erlbaum Associates.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed.). Mahwah N.J.: Lawrence Erlbaum Associates.

Uzun, N. B. ve Gelbal, S. (2017). PISA fen başarı testinin madde yanlılığının kültür ve dil açısından incelenmesi. *Kastamonu Eğitim Dergisi*, *25*(6), 2427–2446.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33(1), 5–20. Retrieved from https://doi.org/10.3102/1076998607302626

van der Linden, W. J. & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A.W. Glas (Eds.), *Elements of adaptive testing*. Springer.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. New developments in psychometrics (s. 207–214). Springer.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed., p. 1–22). Mahwah N.J.: Lawrence Erlbaum Associates.

Wainer, H. & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed.). Mahwah N.J.: Lawrence Erlbaum Associates.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Doctoral Dissertation). Michigan State University.

Wang, S., Haiyan, L., Chang, H.H. & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45–62.

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

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Doctoral Dissertation).

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computer adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weissman, A. (2014). Irt-based multistage testing. In D. Yan, A. A. von-Davier ve C. Lewis (Eds.), *Computerized multistage testing* (s. 153–168). CRC Press; Taylor&Francis Group.

Xiong, X. (2018). A hybrid strategy to construct multistage adaptive tests. *Applied Psychological Measurement*, *42*(8), 630–643. https://doi.org/10.1177/0146621618762739

Yan, D. (2010). *Investigation of optimal design and scoring for adaptive multi-stage testing: A tree-based regression approach* (Master Thesis). Fordham University.

Yan, D., von-Davier, A. A. ve Lewis, C. (2014). Overview of computerized multistage tests. In D. Yan, A. A. von-Davier ve C. Lewis (Eds.), *Computerized multistage testing* (p. 3–20). CRC Press; Taylor&Francis Group.

Yang, L. (2016). *Enhancing item pool utilization when designing multistage computerized adaptive tests* (Doctoral Dissertation). Michigan State University.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*(2), 125–145.

Zenisky, A., Hambleton, R. K. & Luecht, R. M. (2010). Multistage testing: Issues, design, and research. In W. J. van der Linden & C. A.W. Glas (Eds.), *Elements of adaptive testing*. Springer.

Zenisky, A. L. & Hambleton, R. K. (2014). Multistage test designs: Moving research results into practice. In D. Yan, A. A. von-Davier ve C. Lewis (Eds.), *Computerized multistage testing* (p. 21–38). CRC Press; Taylor&Francis Group.

Zheng, Y., Nozawa, Y., Gao, X. & Chang, H. H. (2012). Multistage adaptive testing for a large classification test: Design, heuristic assembly, and comparison with other testing modes. *ACT Research Report Series*, *2012*(6).

Zheng, Y. & Chang, H. H. (2014). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, *39*(2), 104–118. https://doi.org/10.1177/0146621614544519

Zheng, Y., Wang, C., Culbertson, M. J. & Chang, H. H. (2014). Overview of test assembly methods in multistage testing. In D. Yan, A. A. von-Davier & C.

Lewis (Eds.), C*omputerized multistage testing* (p. 87–100). CRC Press; Taylor&Francis Group.

Zieky, M. J. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (p. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden ve C. A.W. Glas (Eds.), *Elements of adaptive testing*. Springer.

Zwick, R. & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A. A. von-Davier & C. Lewis (Eds.), C*omputerized multistage testing*. CRC Press; Taylor&Francis Group.

EK-A Raju İşaretli Alan İndeksi Değerleri

Madde No	İAİ	Madde No	İAİ	Madde No	İAİ
161	0.810	361	0.958	561	0.889
162	0.975	362	0.753	562	0.934
163	0.954	363	0.827	563	0.824
164	0.852	364	0.796	564	0.965
165	0.829	365	0.858	565	0.983
166	0.840	366	0.976	566	0.758
167	0.885	367	0.937	567	0.976
168	0.981	368	0.813	568	0.825
169	0.817	369	0.807	569	0.988
170	0.948	370	0.959	570	0.956
171	0.952	371	0.869	571	0.999
172	0.955	372	0.855	572	0.773
173	0.826	373	0.915	573	0.823
174	0.787	374	0.795	574	0.790
175	0.862	375	0.759	575	0.916
176	0.803	376	0.813	576	0.990
177	0.801	377	0.997	577	0.925
178	0.774	378	0.996	578	0.767
179	0.799	379	0.826	579	0.977
180	0.914	380	0.810	580	0.791
181	0.752	381	0.876	581	0.899
182	0.760	382	0.921	582	0.775
183	0.853	383	0.775	583	0.986
184	0.798	384	0.957	584	0.951
185	0.852	385	0.867	585	0.951
186	0.910	386	0.785	586	0.985
187	0.927	387	0.762	587	0.786
188	0.875	388	0.813	588	0.856
189	0.835	389	0.970	589	0.852
190	0.939	390	0.801	590	0.949
191	0.898	391	0.754	591	0.958
192	0.996	392	0.765	592	0.979
193	0.936	393	0.892	593	0.802
194	0.814	394	0.846	594	0.983
195	0.771	395	0.792	595	0.940
196	0.854	396	0.987	596	0.886
197	0.899	397	0.888	597	0.910
198	0.883	398	0.989	598	0.856
199	0.812	399	0.826	599	0.956
200	0.848	400	0.872	600	0.904

EK-B DMF'li Madde Oranı %10 iken RMSE, Yanlılık ve Korelasyon Değerleri

	RMSE			Yanlılık			Korelasyon		
Test Uzunluğu	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)
10 madde	0.269	0.382	0.389	0.212	0.299	0.307	0.963	0.924	0.921
20 madde	0.192	0.282	0.306	0.151	0.221	0.240	0.982	0.963	0.953
30 madde	0.164	0.246	0.253	0.130	0.193	0.200	0.987	0.974	0.969
40 madde	0.155	0.236	0.238	0.122	0.184	0.187	0.989	0.977	0.974

EK-C DMF'li Madde Oranı %20 iken RMSE, Yanlılık ve Korelasyon Değerleri

	RMSE			Yanlılık			Korelasyon		
Test Uzunluğu	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)
10 madde	0.268	0.408	0.400	0.211	0.318	0.312	0.963	0.913	0.916
20 madde	0.192	0.284	0.320	0.151	0.223	0.252	0.982	0.962	0.949
30 madde	0.164	0.252	0.270	0.130	0.197	0.213	0.987	0.971	0.964
40 madde	0.153	0.242	0.241	0.121	0.190	0.191	0.989	0.975	0.973

EK-Ç DMF'li Madde Oranı %30 iken RMSE, Yanlılık ve Korelasyon Değerleri

	RMSE			Yanlılık			Korelasyon		
Test Uzunluğu	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)	BBT	ÇAT (1-3-3)	ÇAT (1-2-4)
10 madde	0.269	0.448	0.451	0.212	0.348	0.352	0.963	0.894	0.892
20 madde	0.194	0.301	0.303	0.153	0.237	0.238	0.981	0.955	0.954
30 madde	0.166	0.276	0.269	0.131	0.218	0.212	0.987	0.962	0.965
40 madde	0.153	0.246	0.259	0.121	0.194	0.204	0.989	0.970	0.967

EK-D: Etik Komisyonu Onay Bildirimi

Tarih: 31/12/2018 15:37
Sayı: 35853172-300-E.00000379129

T.C.
HACETTEPE ÜNİVERSİTESİ
Rektörlük

Sayı : 35853172-300
Konu : Başak ERDEM KARA Hk.

EĞİTİM BİLİMLERİ ENSTİTÜSÜ MÜDÜRLÜĞÜNE

Enstitünüz Eğitim Bilimleri Anabilim Dalı Ölçme ve Değerlendirme Bilim Dalı tezli doktora öğrencilerinden **Başak ERDEM KARA**'nın **Prof. Dr. Nuri DOĞAN** danışmanlığında yürüttüğü **"Değişen Madde Fonksiyonunun Biresyselleştirilmiş Bilgisayarlı Testler Üzerindeki Etkisi"**başlıklı tez çalışması, Üniversitemiz Senatosu Etik Komisyonunun **18 Aralık 2018** tarihinde yapmış olduğu toplantıda incelenmiş olup, etik açıdan uygun bulunmuştur.

Bilgilerinizi ve gereğini saygılarımla rica ederim.

e-imzalıdır
Prof. Dr. Rahime Meral NOHUTCU
Rektör Yardımcısı

Evrakın elektronik imzalı suretine https://belgedogrulama.hacettepe.edu.tr adresinden 526005bd-397b-4094-84d8-35355bb72f06 kodu ile erişebilirsiniz.
Bu belge 5070 sayılı Elektronik İmza Kanunu'na uygun olarak Güvenli Elektronik İmza ile imzalanmıştır.

Hacettepe Üniversitesi Rektörlük 06100 Sıhhiye-Ankara
Telefon:0 (312) 305 3001-3002 Faks:0 (312) 311 9992 E-posta:yazimd@hacettepe.edu.tr İnternet Adresi: www.hacettepe.edu.tr
Duygu Didem İLERİ

EK-E: Etik Beyanı

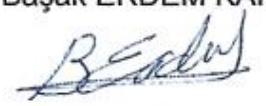
Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında,

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

beyan ederim.

12/11/2019

(İmza)
Başak ERDEM KARA



EK-F: Doktora Tez Çalışması Orijinallik Raporu

13/11/2019

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

Tez Başlığı : Değişen Madde Fonksiyonu Gösteren Madde Oranının Bireyselleştirilmiş Bilgisayarlı ve Çok Aşamalı Testler Üzerindeki Etkisi

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

Rapor Tarihi	Sayfa Sayısı	Karakter Sayısı	Savunma Tarihi	Benzerlik Oranı	Gönderim Numarası
12/11/2019	79	128.062	04/10/2019	%3	1212220538

Uygulanan filtreler:

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

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

Ad Soyadı: Başak ERDEM KARA
Öğrenci No.: N15141802
Ana Bilim Dalı: Eğitim Bilimleri
Programı: Eğitimde Ölçme ve Değerlendirme
Statüsü: ☐ Y.Lisans ☒ Doktora ☐ Bütünleşik Dr.

İmza

DANIŞMAN ONAYI

UYGUNDUR.
(Prof. Dr., Nuri DOĞAN)

EK-G: Dissertation Originality Report

13/11/2019

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

Thesis Title: The Effect of Item Ratio Indicating Differential Item Functioning on Computer Adaptive and Multi Stage Tests

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

Time Submitted	Page Count	Character Count	Date of Thesis Defense	Similarity Index	Submission ID
12/11/2019	79	128.062	04/10/2019	3%	1212220538

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

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

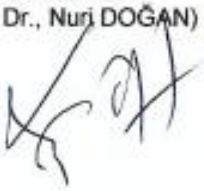
I respectfully submit this for approval.

Name Lastname: Başak ERDEM KARA
Student No.: N15141802
Department: Educational Sciences
Program: Educational Measurement and Evaluation
Status: ☐ Masters ☒ Ph.D. ☐ Integrated Ph.D.

Signature

ADVISOR APPROVAL

APPROVED
(Prof. Dr., Nuri DOĞAN)



EK-Ğ: Yayımlama ve Fikrî Mülkiyet Hakları Beyanı

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

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

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

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

12 /11/2019

(imza)

Başak ERDEM KARA

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

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

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

(3) Madde 7. *1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, tezin yapıldığı kurum tarafından verilir*. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, ilgili kurum ve kuruluşun önerisi ile enstitü veya fakültenin uygun görüşü Üzerine üniversite yönetim kurulu tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
Madde 7.2. *Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir*

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