TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR UYARLAMALI TESTLERDE MADDE YANIT SÜRELERİNİN KULLANIMI

Yusuf Kemal ARSLAN

BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Atilla Halil ELHAN

ANKARA
2021

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR UYARLAMALI TESTLERDE MADDE YANIT SÜRELERİNİN KULLANIMI

Yusuf Kemal ARSLAN

BİYOİSTATİSTİK ANABİLİM DALI
DOKTORA TEZİ

DANIŞMAN
Prof. Dr. Atilla Halil ELHAN

ANKARA
2021

ETİK BEYAN

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Doktora tezi olarak hazırlayıp sunduğum "BİLGİSAYAR UYARLAMALI TESTLERDE MADDE YANIT SÜRELERİNİN KULLANIMI" başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan deneysel çalışma/araştırma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Yukarıda belirtilen hususların doğruluğunu beyan ederim.

Öğrencinin Adı Soyadı: Yusuf Kemal ARSLAN

Tarih:

İmza:

İÇİNDEKİLER

ÖNSÖZ

Bu tez kapsamında, bir psikometrik değerlendirmede veya sınavda kişilerin maddelere verdikleri yanıt sürelerini kullanarak, kişilerin çözümsel veya hızlı yanıt davranışınına sahip olmaları durumunda yetenek düzeylerinin kestiriminde bilgisayar uyarlamalı testlerin performansı benzetim çalışması ile değerlendirilmiştir.

Ankara Üniversitesi'ne ilk geldiğim andan itibaren hiçbir koşulda desteğini benden esirgemeyen ve bilimsel olarak bana ciddi katkıları olan, mesai saati kavramı gözetmeksizin bu pandemik krizde bana her türlü desteği sağlayan kıymetli danışman hocam Prof. Dr. Atilla Halil ELHAN'a,

Tez izleme komitelerimde bulunan ve Ankara'da bulunduğum süre boyunca hiçbir koşulda beni geri çevirmeyip, benden yardımlarını esirgemeyen Prof. Dr. Erdem KARABULUT'a ve Doç. Dr. Beyza DOĞANAY ERDOĞAN'a ve

Savunma sınavımda bulunan ve eğitim sürecimde bana destek olan Doç. Dr. Derya GÖKMEN'e ve Doç. Dr. Can ATEŞ'e,

Bana zor zamanlarımda destek olan ve kıymetli zamanını bana ayıran değerli hocam Doç. Dr. Mergül ÇOLAK'a,

Tez süresince bana yadsınamayacak destekleri olan, dibe vurduğum her anda beni motive edecek bir bakış açısıyla manevi desteğini sürekli üzerimde hissettiğim, özverili ve kıymeli dostum, dönem arkadaşım Arş. Gör. Afra ALKAN'a ve Selen Begüm UZUN'a

Doktora eğitimim boyunca bana birçok katkıda bulunan Ankara Üniversitesi Tıp Fakültesi Biyoistatistik Anabilim Dalı'ndaki tüm hocalarıma, bölüm arkadaşlarıma,

Tüm eğitim sürecimde ve hayatımda bana destek olan, iyi ve kötü günlerimde her zaman yanımda olan ve en büyük teşekkürü hakeden kıymetli AİLEM'e

teşekkürü bir borç bilirim.

SİMGELER VE KISALTMALAR

1PL : Tek Parametreli Lojistik model

2PL : İki Parametreli Lojistik model

3PL : Üç Parametreli Lojistik model

BUT : Bilgisayar Uyarlamalı Test

CAT : Computerized Adaptive Test

ÇD : Çözümsel Davranış

EAP : Expected a Posteriori

ICC : Intraclass Correlation Coefficient

KTT : Klasik Test Teorisi

ML : Ençok olabilirlik

MYT : Madde Yanıt Teorisi

RT : Response Time

SB : Solution Behavior

SEM : Standard Error of Measurement

SH : Standart Hata

SKK : Sınıfiçi Korelasyon Katsayısı

sn : Saniye

θ : Theta, yetenek düzeyi

YS : Yanıt süresi

YSÇ : Yanıt süresi çabası – Response Time Effort

ŞEKİLLER

ÇİZELGELER

1. GİRİŞ

1.1. Araştırmanın Konusu ve Amacı

Günlük yaşantımızda ve bilimsel alanda ölçme önemli bir yer tutar. Ölçme, genel anlamıyla "herhangi bir niteliği gözlemek ve gözlem sonuçlarını sayılarla veya başka sembollerle" ifade etmektir. Bir hastanın boyunun/kilosunun ölçülmesi fiziksel bir ölçme olup "doğrudan ölçme" kapsamına girerken, aynı hastanın özürlülük düzeyinin ölçülmesi ise psikolojik ölçme olup "dolaylı ölçme" kapsamına girmektedir (Turgut,1997). Ölçme araçlarından elde edilen sonuçlar, kişilerin incelenen özellikleri hakkında karar vermede kullanılır.

Ölçek ya da soru bankası geliştirmenin amacı, uygun maddelerden oluşan, güvenilir, geçerli ve değişime duyarlı bir ölçme aracı oluşturmaktır. Bir özellik ölçülürken ölçeği oluşturan tüm maddeler değerlendirmede kullanılır. Bu yüzden ölçülmek istenen özellik en az madde ile en etkin şekilde ölçülmeye çalışılır. Soru bankası ise bir alanda farklı zorluk düzeylerine sahip birçok maddeyi içermelidir. Bu maddeler, aynı alanda değerlendirme yapan farklı ölçeklere ait maddeler olabileceği gibi uzman kişilerce geliştirilen yeni maddeler de olabilir. Madde analiz çalışmaları, gerek ölçek gerekse soru bankası geliştirme sürecinde önemli bir yer tutmaktadır.

Ölçek ya da soru bankası geliştirmede kullanılan iki yaklaşım, klasik test teorisi (classical test theory) ve madde yanıt teorisidir (item response theory). Klasik test teorisi (KTT), bir kişinin bir özelliğe sahip oluş derecesini, ölçme aracının maddelerine verilen cevapların toplamı ile belirler. Modern test teorisi altında yer alan madde yanıt teorisi ise, kişinin bir özelliğe sahip oluş derecesini, kişinin test maddesine verdiği yanıt ile test maddelerine ait parametreler arasındaki ilişkiye dayalı matematiksel bir model olarak açıklar.

Bu tez çalışmasının amacı üç parametreli (3PL) madde yanıt teorisi (MYT) modeliyle, yanıt süresini göz önünde bulunduran Çözümsel Davranış Modelinin (Solution Behavior Model) Bilgisayar Uyarlamalı Test (BUT) performanslarını karşılaştırmaktır. Madde yanıt süresinin modele dahil edilmesinin incelenen özellik düzeyi kestirimi üzerine etkisinin incelenmesi de bu tez çalışmasının ikincil amacı olacaktır. Böylece tez çalışması kapsamında; farklı özellikteki ölçme araçlarıyla uygulamalar yapılarak yanıt sürelerinin modelde yer alma gerekliliği ortaya koyulabilecektir. 3PL modelinin kullanımına ek olarak yanıt sürelerinin yer aldığı modeli yaygın hale getirip üstünlük veya kısıtlılıkları ortaya konulacaktır.

1.2. Madde Yanıt Teorisi ve Klasik Test Teorisi

Ölçme alanında sıklıkla kullanılan iki teoriden biri olan MYT, KTT ile karşılaştırıldığında sağlık alanında elde edilen ölçümlerin belirlenmesi için birçok avantaja sahiptir. KTT'de elde edilen istatistikler, örneğin madde zorluk parametresi (doğru yanıt oranı), madde ayırt edicilik parametresi (düzeltilmiş madde-toplam test korelasyonu) ve güvenilirlik, ölçme aracının uygulandığı gruba bağımlıdır. MYT'de ise parametre tahminleri kullanılan örnekleme bağımlı değildir ve araştırma evrenindeki farklı gruplar için değişmediği varsayılır. Ayrıca, KTT güvenirlik için tek bir tahmin ve standart hata ölçümü verirken; MYT, ölçme aracının kesinliğini, incelenen özellik düzeyi boyunca tahmin eder. KTT'nin diğer bir dezavantajı, kişinin ölçüm aracındaki sonucunun analizde kullanılan maddelere bağımlı olmasıdır. MYT'de ise, kişinin incelenen özelliği, kullanılan maddelerden bağımsızdır. Kişinin incelenen özelliği, her bir maddeye verdiği yanıtlardan hesaplandığı için kestirim skoru, KTT'den elde edilen toplam skordan daha doğru bir kestirimdir (Reeve,2002).

Özellikle psikoloji ve rehabilitasyon alanlarında uygun tedavinin belirlenmesi, izlenmesi ve etkinliğinin değerlendirilmesi amacıyla çeşitli sonuç değerlendirim ölçekleri kullanılmaktadır. Bu ölçeklerin sadece klinik uygulamalarda değil, özellikle gelişmiş ülkelerde daha kapsamlı olarak, toplum içi ve toplumlar arası farklı tedavi programlarının karşılaştırılması, sağlık politikalarının belirlenmesi ve sağlık hizmetleri için kaynak tahsisi gibi alanlarda da kullanımı gündeme gelmiştir. Sonuç

değerlendirimi ve izleniminin en doğru şekilde yapılabilmesi için, bu ölçeklerin kullanıldıkları toplumlara göre din, dil ve sosyokültürel adaptasyonlarının yapılması, geçerlik ve güvenirliklerinin gösterilmesi şarttır. Gerek toplum içinde aynı hastalık grubunda farklı tedavi programları veya farklı hasta gruplarına ait sonuçlarının karşılaştırılması, gerekse toplumlar arası karşılaştırılmalar yapılabilmesi için bu ölçeklerin ulusal ve uluslararası düzeyde standart hale getirilmesi gerekmektedir (Elhan, 2005).

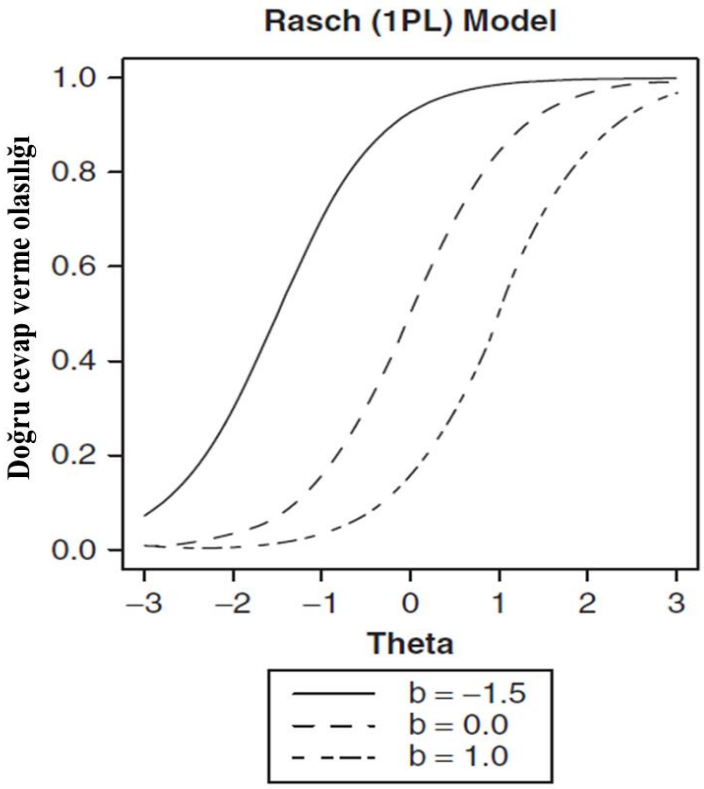
1.3. Madde Yanıt Teorisi Modelleri

1.3.1. Rasch Modeli (1PL)

MYT altında yer alan Rasch modeli tek parametreli bir model olup, bu modelde madde zorluk parametresinin (b_j) yanı sıra kişinin θ düzeyi de bulunmaktadır. Georg Rasch (1960) tarafından iki sonuçlu maddeler için geliştirilen Rasch modelinde, kişinin bir maddeye belirli bir yanıtı verme olasılığı, kişinin θ düzeyi ile madde zorluk parametresi arasındaki farkın lojistik fonksiyonu olarak hesaplanır.

$$P_j(\theta_i) = P(Y_{ij} = 1 \mid \theta_i) = \left(\frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \right) \qquad (1.1)$$

θ, incelenen özellik düzeyini göstermektedir. Sağlık alanında kullanılan ölçeğe ya da soru bankasına bağlı olarak "işlevsellik, özürlülük, tükenmişlik, depresyon ya da yaşam kalitesi düzeyi" olarak adlandırılabilir. Literatürde iki sonuçlu ve çok sonuçlu yanıt kategorileri için geliştirilmiş farklı Rasch modelleri bulunmaktadır (Elhan, 2002). Farklı maddeler için tek parametreli model Şekil 1.1'de gösterilmiştir.

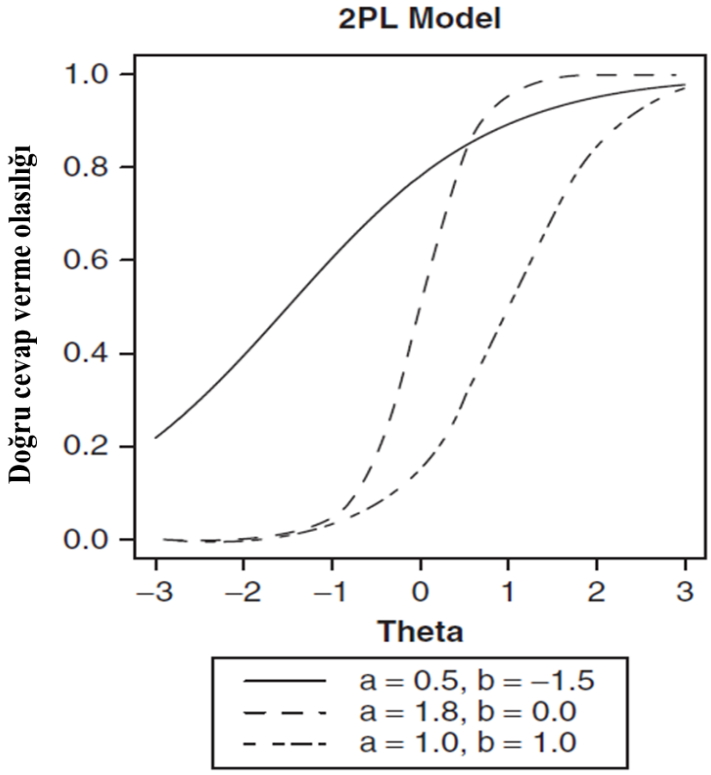


Şekil 1.1. Farklı parametreler için Rasch modeli (Salkind,2010)

1.3.2. 2PL Model

İki parametreli model, Rasch modeline madde ayırt edicilik parametresinin eklendiği modeldir. Bu modelde madde ayırt edicilik parametresi a_j ile gösterilir. Modelde ifade edilen D sabit terimdir (ölçekleme terimi-scale factor). Farklı maddeler için iki parametreli model Şekil 1.2'de gösterilmiştir.

$$P_j(\theta_i) = P(Y_{ij} = 1 \mid \theta_i) = \left(\frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}} \right) \quad (1.2)$$



Şekil 1.2. Farklı parametreler için 2PL model (Salkind,2010)

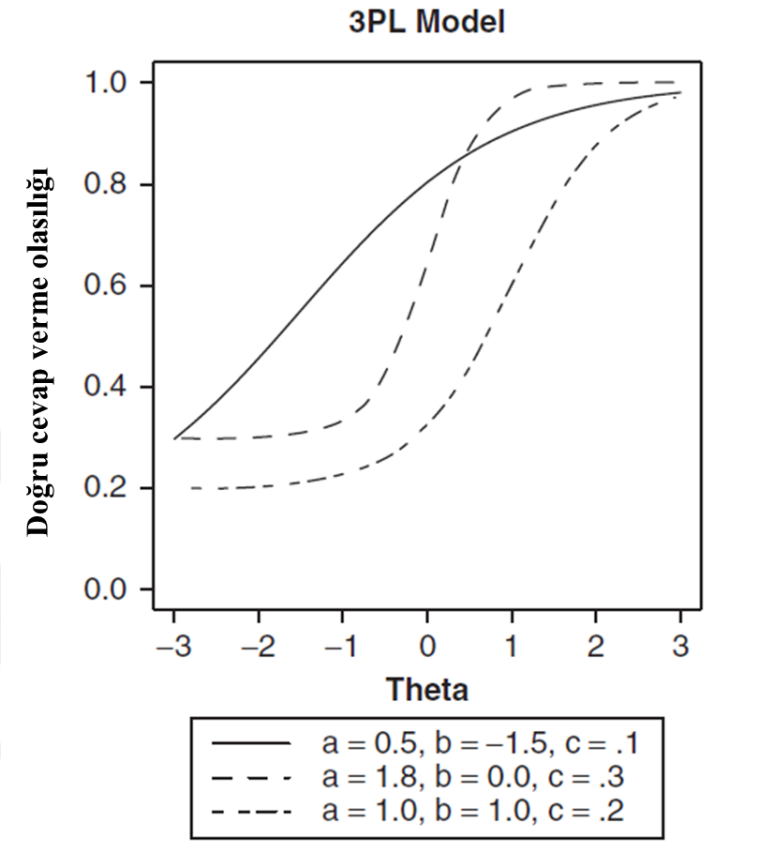
1.3.3. 3PL Model

Birnbaum'un üç parametreli lojistik modeli yaygın olarak kullanılan bir modeldir. θ kişi yetenek düzeyi belliyken bir maddeyi doğru yanıtlama olasılığını veren bu modelde 2PL modeldeki parametrelere ek olarak tahmin parametresi (c_j) (şansa bağlı tahmin – pseudo guessing – alt asimptot) bulunmaktadır.

$$P_j(\theta_i) = P(Y_{ij} = 1 \mid \theta_i) = c_j + (1 - c_j)\left(\frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_i(\theta_i - b_j)}}\right) \tag{1.3}$$

3PL modelin sıklıkla kullanılmasının nedeni, esnek bir model olmasıdır. Fakat literatürde 3PL model kurulurken örneklem genişliği için kesin bir sayı olmamakla beraber yüksek sayıda birime ihtiyaç duyulduğu belirtilmiştir (De Ayala, 2013). Düşük örneklem büyüklüğü ile yapılan çalışmalarda bu kısıtlama parametre

kestirimlerindeki belirsizliği arttırabilir. Farklı maddeler için 3PL model Şekil 1.3'te gösterilmiştir.



Şekil 1.3. Örnek maddeler için 3PL model (Salkind,2010)

1.4. Çözümsel Davranış (Efor Kullanımlı) Modeli

Wise ve Kong (2005), zaman parametresini kullanarak hızlı yanıtlama ve çözüm davranış ayrımını yapabilmek için bir hipotez üretmiş ve bu hipoteze bağlı olarak YS temelli kişi çabasını temsil eden bir indeks oluşturmuşlardır. İki sonuçlu maddelerde belirli bir T eşik değeri için i. kişinin j. maddeye verdiği yanıta ait çözümsel davranış (ÇD);

$$ÇD_{ij} = \begin{cases} 1 & eğer\ YS_{ij} \geq T_j \\ 0 & diğer \end{cases} \quad (1.4)$$

şeklinde ifade edilir. Doğru yanıtlama olasılığına ait model ise aşağıdaki şekilde ifade gösterilir:

$$P_j(\theta_i) = \text{ÇD}_{ij}\ (\text{çözüm davranış modeli}) + (1 - \text{ÇD}_{ij})(\text{hızlı yanıt modeli})$$

Bu denklemden yola çıkarak çözüm davranışı modelinin 3PL MYT modeli olduğu düşünülür ve hızlı yanıt davranışı bir sabit olasılık modeli olarak kabul edilirse model Eşitlik 1.5'teki gibi ifade edilir:

$$P_j(\theta_i) = \left(\text{ÇD}_{ij}\right)\left(c_j + \left(1 - c_j\right)\left(\frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}}\right)\right) + \left(1 - \text{ÇD}_{ij}\right)\left(g_j\right) \qquad (1.5)$$

Bu modelde temel özellik, yanıtlayıcının çabasını içeren bir MYT modeli geliştirmek ve hızlı tahmin etme davranışını göz önünde bulundurmaktır. Modelde bulunan g_j, 1'in o soruda bulunan olası seçenek sayısına bölünmesiyle elde edilir.

1.5. İki Olası Sonuçlu Test İçin Sıralı Yanıt Modeli

Bu tez çalışmasında kullanılan ve eğitim sistemine uygun olduğu düşünülen, 5-seçenekli test sonucunda elde edilen ve olası iki sonuçlu maddeler için (0: Yanlış ve 1: Doğru olmak üzere) yanıt olasılıkları 3PL model ile Eşitlik 1.6 ve 1.7'de verilmiştir. Burada $\boldsymbol{P_{i,1}}$ birikimli yanıt olasılığı fonksiyonudur.

$$P_{i,1}(\theta) = p_i = c_j + \left(1 - c_j\right)\frac{1}{1 + e^{-a_j(\theta - b_j)}} \qquad (1.6)$$

$$P_{i,0}(\theta) = q_i = 1 - P_{i,1}(\theta) = \left(1 - c_j\right)\frac{e^{-a_j(\theta - b_j)}}{1 + e^{-a_j(\theta - b_j)}} \qquad (1.7)$$

Temel fonksiyon;

$$A_{x_i}(\theta) = \frac{\partial}{\partial\theta} log P_{x_i}(\theta) = \left[\frac{1}{P_{x_i}(\theta)}\right]\left[\frac{\partial}{\partial\theta} P_{x_i}(\theta)\right] = \frac{a_j}{1-c_j}\frac{(p_i - c_j)q_i}{p_i} \tag{1.8}$$

Yanıt olasılık fonksiyonun θ'ya göre 2. kısmi türevi;

$$\frac{\partial^2}{\partial\theta^2} P_{x_i}(\theta) = \left(\frac{a_j}{1-c_j}\right)^2 (p_i - c_j)q_i(q_i - p_i + c_j) \tag{1.9}$$

Madde yanıt bilgi fonksiyonu;

$$I_{x_i}(\theta) = -\frac{\partial^2}{\partial\theta^2} log P_{x_i}(\theta) = -\frac{\partial}{\partial\theta} A_{x_i}(\theta)$$

$$I_{x_i}(\theta) = \left[\frac{1}{P_{x_i}(\theta)}\right]^2 \left[\frac{\partial}{\partial\theta} P_{x_i}(\theta)\right]^2 - \left[\frac{1}{P_{x_i}(\theta)}\right]\left[\frac{\partial^2}{\partial\theta^2} P_{x_i}(\theta)\right] \tag{1.10}$$

$$= \left(\frac{a_j}{1-c_j}\right)^2 \left\{\left[\frac{(p_i - c_j)q_i}{p_i}\right]^2 - \frac{(p_i - c_j)q_i(q_i - p_i + c_j)}{p_i}\right\}$$

Madde bilgi fonksiyonu;

$$I_i(\theta) = \sum_{x_i=1}^{k} P_{x_i}(\theta) I_{x_i}(\theta) \tag{1.11}$$

N maddeli bir test için test bilgi fonksiyonu;

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta) \tag{1.12}$$

Log olabilirlik;

$$Olabilirlik = \begin{cases} P(X=0|\theta), & x=0 \\ P(X=1|\theta), & x=1 \end{cases}$$

$$l = \begin{cases} q, & x=0 \\ p, & x=1 \end{cases} \tag{1.13}$$

$$L = log(l) = \begin{cases} \log(q), & x=0 \\ \log(p), & x=1 \end{cases}$$

Doğru yanıt için 1. türev;

$$\frac{\partial}{\partial\theta}p = \frac{a}{1-c}(p-c)q \tag{1.14}$$

Doğru yanıt için 2. türev;

$$\frac{\partial^2}{\partial\theta^2}p = \left(\frac{a}{1-c}\right)^2(p-c)q(q-p+c) \tag{1.15}$$

Yanlış yanıt için 1. türev;

$$\frac{\partial}{\partial\theta}q = -\frac{\partial}{\partial\theta}p = \frac{-a}{1-c}(p-c)q \tag{1.16}$$

Yanlış yanıt için 2. türev;

$$\frac{\partial^2}{\partial\theta^2}q = -\frac{\partial^2}{\partial\theta^2}p = -\left(\frac{a}{1-c}\right)^2(p-c)q(q-p+c) \tag{1.17}$$

Doğru ve yanlış yanıtlar için log-olabilirliğe ait 1. ve 2. türevler;

$$\frac{\partial}{\partial\theta}L=\begin{cases}\frac{\partial}{\partial\theta}log(q)=\frac{\frac{\partial}{\partial\theta}q}{q}=\frac{-p'}{q}, & x=0\\ \frac{\partial}{\partial\theta}log(p)=\frac{\frac{\partial}{\partial\theta}p}{p}=\frac{p'}{p}, & x=1\end{cases} \tag{1.18}$$

$$=\begin{cases}\frac{-a}{1-c}\frac{(p-c)q}{q}, & x=0\\ \frac{a}{1-c}\frac{(p-c)q}{p}, & x=1\end{cases}$$

$$\frac{\partial^2}{\partial\theta^2}L=\begin{cases}\frac{\partial^2}{\partial\theta^2}log(q)=-\frac{p''}{q}-\left(\frac{p'}{q}\right)^2, & x=0\\ \frac{\partial^2}{\partial\theta^2}log(p)=\frac{p''}{p}-\left(\frac{p'}{p}\right)^2, & x=1\end{cases} \tag{1.19}$$

$$=\begin{cases}-\left(\frac{a}{1-c}\right)^2\frac{(p-c)q(q-p+c)}{q}-\left(\frac{a}{1-c}\right)^2\left(\frac{(p-c)q}{q}\right)^2, & x=0\\ \left(\frac{a}{1-c}\right)^2\frac{(p-c)q(q-p+c)}{p}-\left(\frac{a}{1-c}\right)^2\left(\frac{(p-c)q}{p}\right)^2, & x=1\end{cases}$$

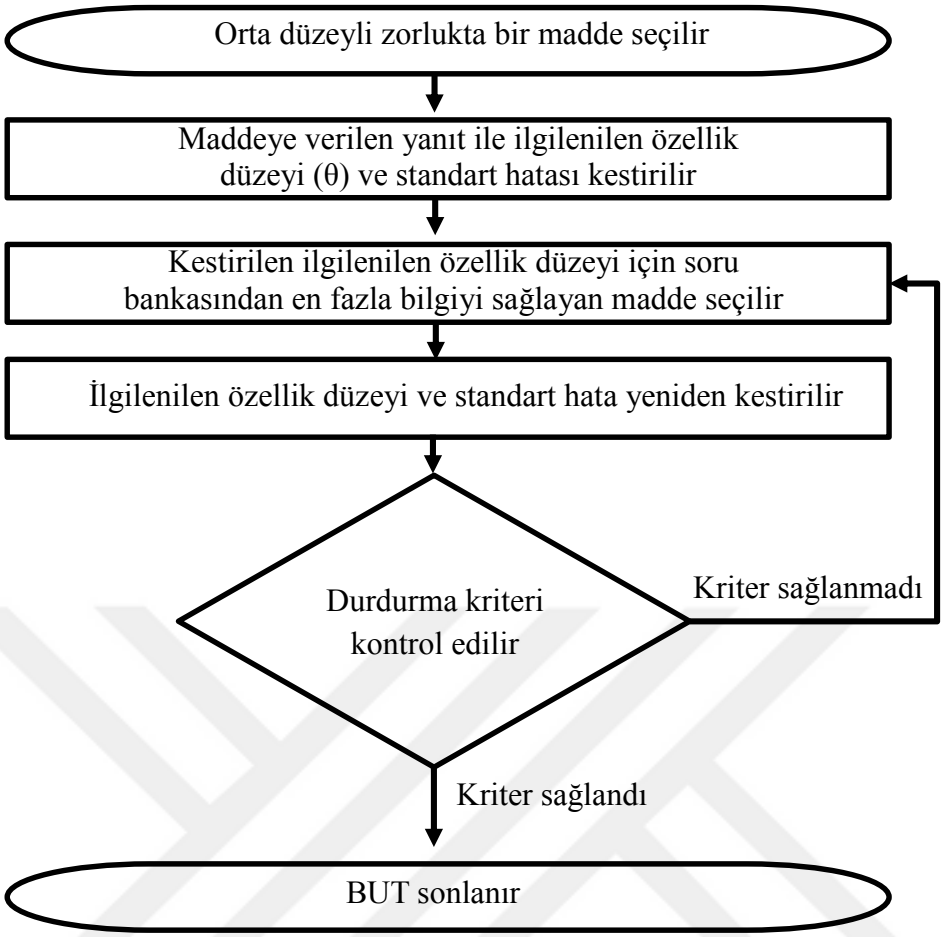
şeklinde ifade edilir.

1.6. Bilgisayar Uyarlamalı Test

Bilgisayar uyarlamalı testler, ilgilenilen özelliğin ölçümünde sıklıkla kullanılan etkili bir yöntemdir. Katılımcının başlangıçta sorulan orta zorluk düzeyli maddeye verdiği yanıt, daha sonra gelecek olan maddelerin sırasında belirleyici olmaktadır. Katılımcı, başlangıçta sorulan orta zorluk düzeyine sahip maddeyi doğru yanıtlandıysa, takip eden madde daha zor, yanlış yanıtlandıysa takip eden madde

daha kolay olur. BUT yaklaşımıyla, katılımcının ölçülen özelliğini (yeteneğini, θ) kestirmede çok etkili olmayan uç noktalardaki zorluk düzeyine sahip maddeler (çok zor veya çok kolay) yerine, katılımcının yetenek düzeyine uygun olan maddelerin seçilmesi sağlanır. Kişinin uygulanan orta zorluk düzeyli maddeye verdiği yanıt sonrasında ilgilenilen özellik kestirilir ($\hat{\theta}$) ve diğer maddelerden elde edilebilecek bilgi düzeyleri hesaplanır. En yüksek bilgiyi sağlayan madde katılımcıya uygulanır ve aynı işlem sürecin durması için belirlenen kriter sağlanana kadar devam eder. Bu durdurma kriterleri; süre, madde sayısı, $\hat{\theta}$'daki değişim veya $\hat{\theta}$'ya ait standart hatanın belirli bir eşik değerin altına inmesi olabilir.

BUT'un temel amacı kişinin yetenek düzeyine uygun maddeleri seçmektir. Kişiye daha az sayıda madde uygulanarak kestirim yapılır. BUT algoritması Şekil 1.4'te verilen akış şemasına uygun olarak çalışır.



Şekil 1.4. BUT akış şeması

BUT ile θ kestirilirken klasik istatistik yöntemler ve Bayesci yaklaşım kullanılabilir. Sıklıkla tercih edilen kestirim yöntemleri en çok olabilirlik (Maximum Likelihood) ve sonsal beklentidir (Expectation a Posteriori) (Si ve Schumacker, 2004). θ kestirimi dışında soru seçimi için de bu yaklaşımlar kullanılabilir.

1.6.1. BUT Uygulamasının Avantajları ve Dezavantajları

BUT'ta birçok katılımcı için eşit derecede kesinlikte yetenek düzeyi elde edilebilir (Weiner ve ark., 2000). BUT uyarlanabilir bir test olması sebebiyle klasik test tekniğine göre uygulanan madde sayıları %50 azalabilir ve klasik test tekniğine yakın kesinlikte yetenek düzeyi kestirimleri yapılabilir (Weiss ve Kingsbury, 1984). BUT uygulaması sınava giren kişi için bir zaman tasarrufu sağlar; kişiler çok zor veya önemsiz derecede kolay olan maddeleri yanıtlamak için zamanı boşa

harcamazlar. Büyük hedef kitleler ile yürütülen bilimsel ve araştırmaya dayalı alanlarda BUT uygulaması tercih edilmektedir. BUT, engellilik veya hastalıkların erken başlangıcını yakalamak için de kullanılabilir. Herhangi bir bilgisayar tabanlı test gibi, uyarlamalı testlerde de testin tamamlanmasından hemen sonra sonuçlar elde edilebilir. Uyarlamalı testlerde, madde seçim algoritmasına bağlı olarak, bazı maddelere maruziyet azaltılabilir; çünkü sınava giren kişilere klasik testteki gibi tüm maddeleri uygulamak yerine farklı maddeler uygulanır. Bununla birlikte, bazı durumlarda maruziyet artabilir (testin başında kişilere yöneltilen orta veya orta/kolay düzeyli maddeler).

Uyarlanabilir testlerde, birkaç maddenin aşırı kullanımını (maruz kalma) önlemek için kontrol algoritmaları olsa da, yeteneğe bağlı maruz kalma genellikle kontrol edilmez (Weiner ve ark., 2000). Yani aynı yeteneğe sahip katılımcılar için bazı maddelerin testlerde çok kullanılması yaygındır. Bu, ciddi bir güvenlik sorunudur. Aslında, tamamen rastgele bir sınav en güvenli (ama aynı zamanda en az verimli) olandır.

Uyarlanabilir testler geçmiş maddelerin gözden geçirilmesine genellikle izin vermez, çünkü bir kişiyi yanlış cevap verdikten sonra daha kolay maddelere yöneltme eğilimindedir. Sınava giren kişiler sıklıkla maddeleri yeniden gözden geçirememe konusunda şikâyet ederler (Rudner, 1998).

Literatürde araştırmacılar, BUT kullanımıyla ilgili bir takım avantajdan bahsetmişlerdir. Bunlardan ilki test yönetiminin esnek olmasıdır. Sınav sonuçları hemen elde edilebilir ve bu durum sınava girenleri motive edebilir (Rudner, 1998). Testteki maddeler katılımcının seviyesine uygun olduğu için kaygı düzeyi azaltılabilir (Rezaie ve Golshan, 2015).

BUT'ların kullanımı, test hazırlama ve işaretleme için gereken süreyi azaltır. Bu da sonuçların tutarlılığını artırabilir (Callear ve King 1997). BUT'un uygulanması kâğıt ve kalem (paper-pen, klasik test) testlerine göre zaman tasarrufu sağlar. Çünkü BUT'ta kullanılan madde sayısı ve bu soruları cevaplamak için gereken süre,

geleneksel kâğıt ve kalem testlerine göre daha azdır (Madsen, 1991; Wainer, 1990). Her test, sınava giren her kişinin seviyesine göre uyarlandığından, geleneksel testlere kıyasla BUT'tan daha fazla bilgi toplanabilir (Young ve ark. 1996).

Diğer test türleri gibi BUT'un da kendine özgü dezavantajları vardır. Bu nedenle, testi uygulayan araştırmacılar bu kısıtlamaların farkında olmalıdır. Literatürde araştırmacılar, BUT kullanımıyla ilgili bir takım dezavantajdan bahsetmişlerdir. BUT'lar tüm madde türleri için kullanılamayan MYT modeline dayanmaktadır. Bu nedenle, açık uçlu sorulara ve kolayca kalibre edilemeyen maddelere uygulanamazlar (Rudner, 1998). Buna ek olarak diğer önemli bir dezavantaj, sınava giren kişinin cevapladığı sorulara geri dönmesine ve cevapları değiştirmesine izin verilmemesidir, çünkü sonraki maddeler önceki maddelere göre seçilir (Rudner, 1998). Ayrıca, madde kalibrasyonu, bir BUT'un başarısını etkileyen önemli faktördür. Maddeler, zorluk/yetenek ölçeğine uygun şekilde kalibre edilmemişse, uygulanan test ne geçerli ne de güvenilir olacaktır. Bazı araştırmacılara göre tek boyutluluk BUT'un bir dezavantajıdır (Madsen, 1991). BUT genel olarak MYT veya gizli özellik (latent trait) modellerine dayanmaktadır. Bu modellerin en önemli varsayımlarından biri tek boyutluluktur. Tek boyutluluk, tüm test maddelerinin tek bir özelliği başka özelliklere karıştırmadan ölçmesi anlamına gelir (Rezaie ve Golshan, 2015). Bu varsayım testler için her zaman sağlanmalıdır.

BUT'ta, bazı maddeler diğerlerine göre daha sık seçilebilir ve bu maddeler ezberlenip diğer sınava giren kişilere aktarılabilir (Wainer, 1990). Dolayısyla, güvenlik BUT'ların sınırlamalarından biri olabilir. Bu nedenle, sınava girenlerin bilgilerini test etmek için geniş bir madde havuzu kullanılmalıdır.

1.7. Madde Yanıt Süresi

BUT yönteminde en yaygın olarak kullanılan modellerden birisi 3PL MYT modelidir. Son zamanlarda maddelere verilen yanıt için geçen sürenin de göz önünde bulundurulduğu modeller önerilmiştir. Bu modellerin BUT performanslarına ilişkin çalışmalar çok yenidir. Bu modellere ilişkin literatür bilgisi aşağıda verilmiştir.

Kişilerin sınav esnasında gereken çabayı sarf edip etmediklerini belirlemek genellikle mümkün olmamaktadır. Dolayısıyla elde edilen test puanlarının geçerliği ve yorumlanması karmaşıklaşmaktadır. Bu durumu çözebilmek adına madde yanıt süresi (YS) ölçümünü temel alan BUT uygulamaları kullanılabilir. Yanıt süresi çabası (response time effort, YSÇ) olarak adlandırılan bu ölçüm, bir madde uygulandığında, motivasyonu düşük olan yanıtlayıcıların, çok hızlı bir şekilde cevap vereceği (maddeyi/soruyu okumadan veya tam olarak düşünmeden) hipotezine dayanmaktadır. Literatürde YSÇ skorlarının psikometrik özellikleri ampirik olarak incelenmiş ve skorların güvenirliğine ve geçerliğine ait destekleyici kanıtlar bulunmuştur. Fakat YSÇ ölçümünün potansiyel uygulamaları ve sonuçları tartışmaya açıktır (Wise ve Kong, 2005).

Test uygulayarak bir kişinin yeterliliği ölçülmek istendiğinde, kişinin ne bildiğini ve/veya neler yapabileceğini kesin olarak göstermeye çalışacağı varsayılmaktadır. Bununla birlikte, kişinin test sonucuna dayanarak yapılan çıkarımın geçerliliği, kişinin test boyunca gösterdiği çaba miktarına bağlıdır. Kişi yeterli çaba göstermediğinde gerçek yeterlilik seviyesini temsil etmeyen bir test puanı elde edilir. Yani düşük çaba, test puanının geçerliliği için ciddi bir potansiyel tehdit oluşturur (Wise ve Kong, 2005).

MYT çerçevesinde birçok çalışma, kaydedilen YS'lerin; test tasarımını, BUT'ta madde seçimini, madde kalibrasyonunu ve anormal yanıt tespitini iyileştirmeye yardımcı olabileceğini göstermiştir (Wise ve Kong, 2005; Wise ve DeMars, 2006).

Bazı çalışmalarda yeterlik test performansı ile kişinin çabası arasında ilişki olup olmadığı araştırılmıştır ve çalışmalarda sonuçlar tutarlı bulunmuştur. Düşük motivasyona sahip kişilerin test performanslarının, yüksek motivasyona sahip akranlarına göre daha düşük olduğu sonucuna ulaşılmıştır. Wise ve Demars (2006) halihazırda 15 araştırmayı sentezleyen bir çalışma yürüttüklerini ve hesapladıkları ortalama etki büyüklüğüne göre iki grup arasındaki bu farklılığın test uygulayıcılar için de anlamlı sonuç oluşturacağını bildirmişlerdir. Ancak testte sarf edilen çabanın test sonrasında kişisel bildirim (self-report) olarak elde edilmesi bir handikapa neden

olmaktadır. Kişilerin verdiği geri bildirimlerin doğruluğu belirsizdir. Örneğin iyi motive olmuş kişiler öz değerlendirme yaparken bilgi düzeylerinin düşüklüğünü bildirmek yerine fazla çaba göstermediklerini bildirme eğilimindedir. Diğer bir sıkıntılı durum ise, gerçekten çaba sarf ettiğini bildiren kişilerin ellerinden gelenin en iyisini yapmaması veya testi ciddiye almamasıdır (Wise ve DeMars, 2006).

Bu durumla ilgili sorunu çözebilmek ve kişinin çabasını tespit edebilmek için teorik bir ölçüm modeli kullanılarak kişisel madde yanıt deseni incelenebilir. Anormal yanıt desenini tanımlayabilmek adına kişi bazlı uyum istatistikleri (person-fit statistics) geliştirilmiştir. Anormal yanıt davranışlarından birisi rastgele cevap vermedir ve motivasyonu düşük olan kişiler maddelere hızlı bir şekilde cevap verebilmektedir (Wise ve DeMars, 2006).

Kişi bazlı uyum istatistiklerinin kullanımında iki sakınca bulunmaktadır. Bunlardan ilki bu uyum istatistiklerinin anormal davranış sayısından etkilenmesidir. Bu uyum istatistikleri rastgele cevap vermeye ek olarak, kopya çekme, şansa bağlı tahmin etme gibi durumları belirlemede de kullanılmaktadır. Ayrıca bilişsel tanıda kişilerin kavram yanılgılarını incelemeye yardımcı olmak içinde kullanılmaktadır. İkinci sakınca ise, uyum istatistiklerinin anormal yanıt davranışıyla ilgili genel bir fikir vermesi ve test süresince harcanan çabaya karşı duyarlı olmamasıdır (Wise ve DeMars, 2006).

Kişisel bildirim ile ilgili bu durumu ortadan kaldırabilmek için maddelere/sorulara verilen yanıtları toplamanın yanı sıra her madde için harcanan çabanın (sürenin) kaydedilmesi gerekmektedir. Bu işlemi gerçekleştirebilmek için BUT tercih edilebilir. YS'nin objektif olarak hesaplanmasında BUT tercih edilmesinin sebeplerinden ilki sürelerin yanıtlayıcının dikkatini çekmeyecek ve performasını etkilemeyecek şekilde elde edilmesidir. İkincisi ise yanıtlayıcının davranışıyla ilgili doğrudan gözlem sağlayacağından, kişinin motivasyon süreçlerinden veya kişinin duygularından (öz-bildirim) etkilenmeyecek olmasıdır (Wise ve DeMars, 2006).

YS, ölçme alanında çalışan birçok araştırmacı için ilgi çeken bir ölçüttür. Literatürde YS'nin daha doğru yeterlik düzeyi tahminlerini elde etmede, uygun zaman sınırlarını belirlemede, çalışma hızı sorunları ile ilgili (Schnipke ve Scrams, 1997; Scrams ve Schnipke, 1998) ve sıra dışı test performansının belirlenmesinde (Wise ve Kingsbury, 1994) kullanımı ile ilgili çalışmalar mevcuttur. Ayrıca YS, madde parametrelerini daha doğru şekilde elde edilmesi için verileri filtrelemede de kullanılmıştır (Wise ve DeMars, 2006; Schnipke, 1999).

1.7.1. Madde Yanıt Süresine Ait Eşik Değerin Belirlenmesi

Literatürde madde yanıt sürelerine ait eşik değerleri belirlemek için yapılan birçok çalışma mevcuttur. Bu çalışmalar incelenmiş ve eşik değer belirleme yöntemleri sunulmuştur.

1.7.2. Ortak Eşik

Madde uzunluğuna ya da okuma hızına bağlı bir eşik değer belirlenmez. Genel olarak tüm maddeler için madde yanıt süresi eşik değeri 3 saniye olarak kabul edilir. Pratikte tüm maddeler için ortak bir eşik değer kullanımı uygulanması en kolay yöntemdir.

1.7.3. Normative Yöntem

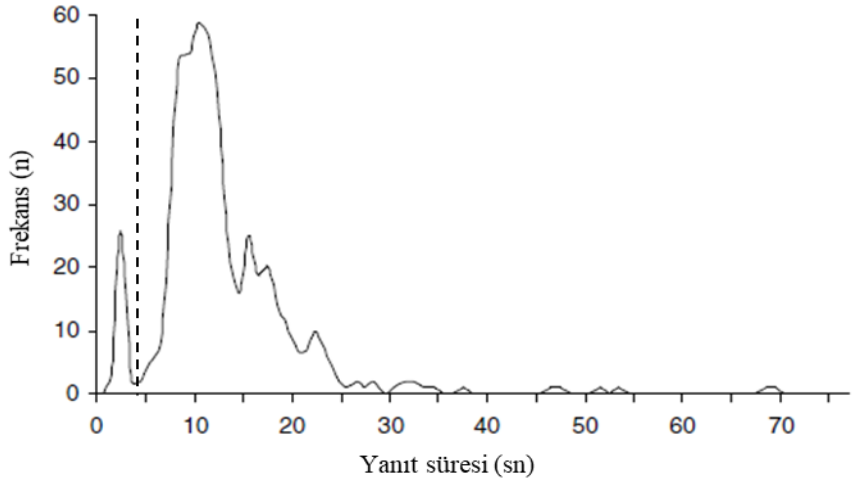
Katılımcıların yanıt verme sürelerinin her madde için ortalaması hesaplanır ve bu ortalamanın %10'u eşik değer olarak kabul edilir. Örneğin, bir maddeye verilen YS ortalaması 40 saniye kabul edilirse eşik değer 4 saniye olacaktır, fakat bu şekilde hesaplanan eşik değer en fazla 10 saniye olabilir. On saniyenin üzerinde eşik değer kullanarak yapılan "hızlı tahmin etme (rapid guessing)" tanımlaması hatalı sonuçlara yol açabilir (Wise ve Ma, 2012).

1.7.4. Okuma ve Karakter Sayısı

Her bir madde için harcanan sürenin maddenin fiziksel özelliği (karakter sayısı, gramatik yapı, içerik) ile ilişkili olduğu düşünülmektedir. Bu nedenle, maddelerin özelliklerine dayanarak her bir madde için bir eşik belirlenebilir. Madde uzunluğu (karakter sayısı) ve şekil, resim ya da başka bir okuma materyali içermesi eşik değer belirlemede kullanılır. Wise ve Kong (2005) çalışmalarında YS için eğer madde 200 karakterden kısaysa 3 saniyelik bir eşik; madde 1.000 karakterden uzunsa veya şekil, resim ya da başka bir okuma materyali içeriyorsa 10 saniyelik bir eşik, kalan durumlar için ise 5 sn'lik bir eşik kullanmıştır.

1.7.5. Görsel Tanımlama

Bu yöntemde, tüm maddelerin YS dağılımları madde bazında incelenerek eşik değer belirlenir (Şekil 1.5). Genellikle YS bimodal (iki-modlu) dağılıma sahiptir. Ortanca YS'den çok önce meydana gelen bir frekans artışı olmaktadır. Her madde için değerlendiriciler YS'yi görsel olarak inceleyebilir ve eşik değer belirleyebilirler (Wise, 2006).



Şekil 1.5. Yanıt süresine ait eşik değerin görsel olarak belirlenmesi (Wise, 2006, syf.22)

1.7.6. İki Aşamalı Karma Model

İki aşamalı karma model, biri hızlı yanıt verenlerin diğeri daha yavaş yanıt verenlerin dağılımını temsil eden bimodal dağılımdan oluşmaktadır. Schnipke ve Scrams, hızlı tahmin etme davranışı ve çözümsel davranışın iki ayrı tepki stratejisi olduğunu çalışmalarıyla desteklemiş ve çözümsel davranış endeksinin hesaplanmasında kullanılacak zaman eşiğini iki dağılımın kesişim noktası olarak belirlemiştir (Schnipke ve Scrams, 1997).

1.8. Kestirim Yöntemleri ve Bilgi Kriteri

1.8.1. En Çok Olabilirlik (Maximum Likelihood-ML) Kestirimi

Yetenek düzeyi belirlenmek istenen bir kişinin başlangıç yetenek düzeyi bilindiğinde farklı sorulara verdiği yanıtlar bağımsız olarak düşünülerek belirli bir yanıt deseninin ortaya çıkma olasılığı θ'nın olabilirliği olarak adlandırılır. Olabilirlik bu yanıt desenindeki tüm yanıt olasılıklarının çarpımı ile hesaplanır. Yetenek düzeyine ait ML kestirimi olabilirliği maksimum yapan değeri bularak yapılır. Bu değeri bulmak için olabilirlik fonksiyonunun θ'ya göre 1. kısmi türevi sıfıra eşitlenir. Olabilirlik fonksiyonu çarpımsal yapıda olduğundan işlem kolaylığı için logaritması alınır. ML kestirimi yapılırken iteratif yöntemlerden faydalanılabilir. Bu yöntemlerden birisi Newton-Raphson'dır ve başlangıçtaki yetenek düzeyi kullanılarak işlem yapılır. Bu adımsal yapıda bir kişinin başlangıç θ'sından, log-olabilirliğin 1. türevinin 2. türevine oranı çıkartılır ve son teta kestirilir. Bu iteratif süreç teta kestirimleri arasındaki fark önemsenmeyecek boyuta (<0,0001) gelene kadar devam eder ve θ'ya ait ML kestirimi yapılmış olur.

1.8.2. Sonsal Beklenti (Expected a Posteriori - EAP) Kestirimi

Sonsal beklenti kestirimi ML kestiriminin aksine iteratif bir süreç olmayıp klasik istatistikçi yaklaşıma alternatiftir. Bayesci istatistik çerçevesinde geliştirilen

bu yaklaşım madde parametrelerinin bilindiği koşulda kullanılır. Madde parametrelerinin ve yanıt deseninin bilindiği durumda θ'ya ait önsel bilgi ile yanıt desenine ait olabilirlik çarpılarak sonsal bilgi elde edilir. Yetenek düzeyine ait önsel bilgi normal dağılımdan veya beta dağılımından seçilebilir. EAP kestirimi, sonsal dağılımın beklenen değerine eşit olmaktadır. Kestirime ait standart hata ise sonsal dağılıma ait standart sapma olacaktır. Belirli yanıt deseni için ölçümün koşullu standart hatasıdır (SEM).

1.8.3. Fisher'in Maksimum Bilgi Kriteri

ML yöntemi ile yapılan hesaplamalarda bulunan varyans bilgi test fonksiyonu olarak da kullanılır. Bilgisayar uyarlamalı testlerde bir sonraki maddenin belirlenmesi için bu bilgi önemlidir. ML sonucunda kestirilen son-teta ve madde havuzundaki maddelere ait parametreler kullanılarak uygulanmamış maddelerin sağlayacağı bilgi düzeyleri kestirilir ve en yüksek bilgiyi sağlayan madde sıradaki yeni madde olarak kişiye yöneltilir.

2. GEREÇ ve YÖNTEM

Bu tez çalışması kapsamında, R dili (ver. 3.6.2) ve RStudio Software (ver. 1.2.5033) yazılımı kullanılarak benzetim çalışması gerçekleştirilmiştir. Kişi sayısı 1000; soru bankasındaki madde sayısı 50, 100 ve 250; BUT durdurma kriteri için standart hata <0,3 ve <0,5 olacak şeklinde toplam 6 senaryo için veri setleri türetilmiştir. Her senaryo için 1000 tekrar uygulanmıştır. Benzetim çalışmasında Firestar (ver. 1.3.2) programı yardımıyla oluşturulan R kodları (Choi, 2009), 3PL model ve ÇD modeline uyarlanmıştır. Bilgi fonksiyonları için uyarlanan kodların doğruluğu "*irtoys"* paketi aracılığıyla kontrol edilmiştir (Partchev, 2016).

Her bir senaryoda, BUT 3PL, BUT ÇD ve tüm maddelerin uygulandığı standart 3PL model ile ÇD modelinden elde edilen θ kestirimleri kaydedilmiştir. BUT performansını belirlemek üzere BUT kestirimlerinin yanlılıkları, standart hataları, uygulanan madde sayısı ortalaması ve standart sapması ile soru bankasındaki maddelerin uygulanma frekansları hesaplanmıştır. Ayrıca kestirimlerin gerçek θ değerleri ile uyumları sınıfiçi korelasyon katsayısı (SKK) ile (ICC(A,1) veya ICC(2,1)) değerlendirilmiştir. SKK "*irr"* paketindeki "*icc()*" fonksiyonu kullanılarak hesaplanmıştır (Gamer, 2012). Yöntemler arasındaki uyum değerlendirilirken Bland-Altman grafikleri "*BlandAltmanLeh*" paketi yardımıyla çizdirilmiştir (Lehnert, 2015). Kestirim yanlılıkları ve standart hataları eşitlik 2.1 ve eşitlik 2.2 yardımıyla elde edilmiştir.

$$\text{Yanlılık} = \frac{\sum_{r=1}^{1000}\left(\hat{\theta} - \theta\right)}{1000} \tag{2.1}$$

$$\text{STD HATA} = \frac{\sum_{r=1}^{1000} SH(\hat{\theta})}{1000} \tag{2.2}$$

2.1. Madde Parametrelerinin Türetilmesi

Oluşturulan soru bankasındaki maddelerin zorluk parametreleri (b_j) için 0 ortalama ve 1,5 standart sapmalı normal dağılım (N(0, 1,5)); ayırt edicilik parametreleri (a_j) için [0, 2] aralığında tekdüze dağılım (U(0,2)) kullanılmıştır. Üç parametreli modelde yer alan ve alt asimptot olarak da isimlendirilen şansa bağlı tahmin parametresi (c_j) için ise [0,2 , 0,5] aralığında tekdüze dağılım (U(0,2 , 0,5)) kullanılmıştır.

2.2. Kişi Yetenek Düzeylerinin Türetilmesi

Gerçek θ düzeyleri N(0, 1,5) dağılımdan türetilmiştir. EAP kestirimleri için [-4, 4] aralığında 0,1 artışla 81 adet θ oluşturulmuştur. Böylece θ'lar, b'lerle benzer dağılım göstermiştir.

2.3. Yanıt Süresinin Türetilmesi

YS, her bir senaryoda kişilerin %10'u hızlı yanıt davranışı, %90'u çözümsel davranış gösterecek şekilde b'lerle ilişkili olarak türetilmiştir.

YS, hızlı yanıt davranışı sergileyen kişiler için b'lerle negatif orta-yüksek düzeyli (r=-0,6) korelasyona, çözümsel davranış sergileyecek kişiler için ise pozitif orta-yüksek düzeyli (r= 0,6) korelasyona sahip olacak şekilde sırasıyla N(6,4) ve N(60,40) dağılımlarından türetilmiştir.

Türetilen yanıt sürelerine ait madde bazlı eşik değer belirlenirken, o maddeye verilen yanıt sürelerinin ortalaması alınmış ve ortalama değerin %10 u yanıt süresine ait eşik değer olarak belirlenmiştir. Bu eşik değer 10 saniyenin üzerinde olan durumlarda 10 sn olarak kabul edilmiştir.

2.4. Yanıt Deseninin Oluşturulması

Türetilen maddeler kullanılarak 3PL model ile kişilerin sorulara verdiği yanıt olasılıkları hesaplanmıştır ve bu olasılıklar rastgele oluşturulan olasılıklar ile (0-1 arasında) karşılaştırılarak yanıt deseni oluşturulmuştur. Eğer hesaplanan olasılık rastgele türetilen olasılıktan büyük ise kişi soruyu doğru yanıtlamış, değil ise kişi soruyu yanlış yanıtlamış kabul edilmiştir. Sonuç olarak 0 veya 1'den oluşan yanıt deseni elde edilmiştir.

2.5. Madde Parametrelerinin Kestirilmesi

Rutin sınav uygulamalarında, araştırmacının elinde yalnızca 0 veya 1'lerden oluşan ve kişilerin soruları yanlış veya doğru yanıtladığı bilgisini içeren bir yanıt deseni elde edilmektedir. Benzetim çalışmasında oluşturulan yanıt deseninden madde parametrelerine ait kestirimler yapılmıştır. Kestirimler yapılırken "*ltm*" paketindeki "*tpm()*" fonksiyonu kullanılmıştır.

2.6. BUT Uygulamasında Soru Seçimi

BUT uygulaması, tüm kişilere orta düzeyli madde sorularak başlatılmıştır. 3PL modele ait BUT için bir sonraki madde, en yüksek bilgiyi veren madde olarak seçilip süreç bu şekilde durdurma kriteri sağlanana kadar devam ettirilmiştir.

Yanıt süresini içeren BUT ÇD için kişiye sorulacak sorularda benzer uygulama yapılmıştır. Her adımda θ'lar ile madde bilgisi hesaplanırken şansa bağlı tahmin olasılığı kullanılmamıştır. Çünkü yanıt olasılığı 0,2 olduğunda madde bilgisi 0 olacağından ilgili maddenin sorulması mümkün değildir. Dolayısıyla YS eşik sürenin altında olan bir maddenin bile kişiye yönlendirilmesi sağlanmıştır.

2.7. Yetenek Düzeyi Kestirimi

İlk sorunun seçilmesi için başlangıç θ, 0 olarak alınmıştır. Newton-Raphson iterasyonu için durdurma kriteri son $\hat{\theta}$ ile bir önceki adımdan elde edilen $\hat{\theta}$ arasındaki farkın 0,0001'den küçük olması şeklinde belirlenmiştir. Newton-Raphson iterasyonu sırasında yakınsama sorunu yaşandığında kestirimin ±4 sınırlarını aşması durumunda $\hat{\theta}$ bu sınır değerler olarak alınmıştır.

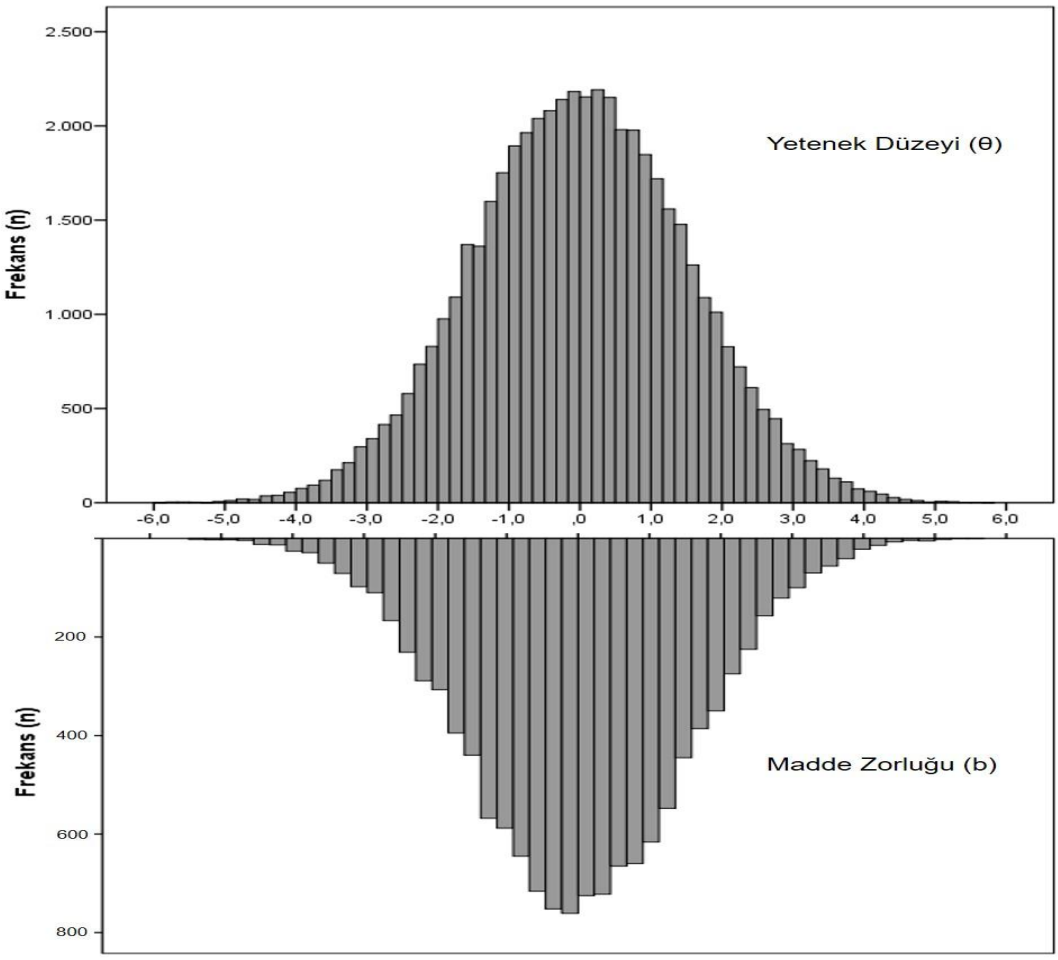
BUT ÇD'de uygulanan ilk maddenin hızlı yanıt davranışı sergilenen bir madde olması durumunda yanıt olasılığı 0,2 olduğundan $\hat{\theta}$ hesabında kullanılan kısmi türevler 0 olmakta ve 0/0 belirlisizliği gerçekleşmektedir. Bu yüzden uygulanan ilk madde hızlı yanıt davranışı sergilenen bir madde olduğunda maddenin doğru cevaplanma olasılığı 0,2 yerine 0,201 olarak alınmıştır.

Yetenek düzeyi kestirimlerinde hibrit yöntem kullanılmıştır (EAP+ML). Yanıt deseni değişmeyen durumlarda EAP kestirimi yapılmış, desen değişikliklerinde ML kestirimiyle işlem devam etmiştir. Maksimum iterasyon sayısı (50) tamamlandığında ve yakınsama gerçekleşmediğinde uç değerler -4 ve +4 kullanılarak bir sonraki adıma geçilmiştir.

3. BULGULAR

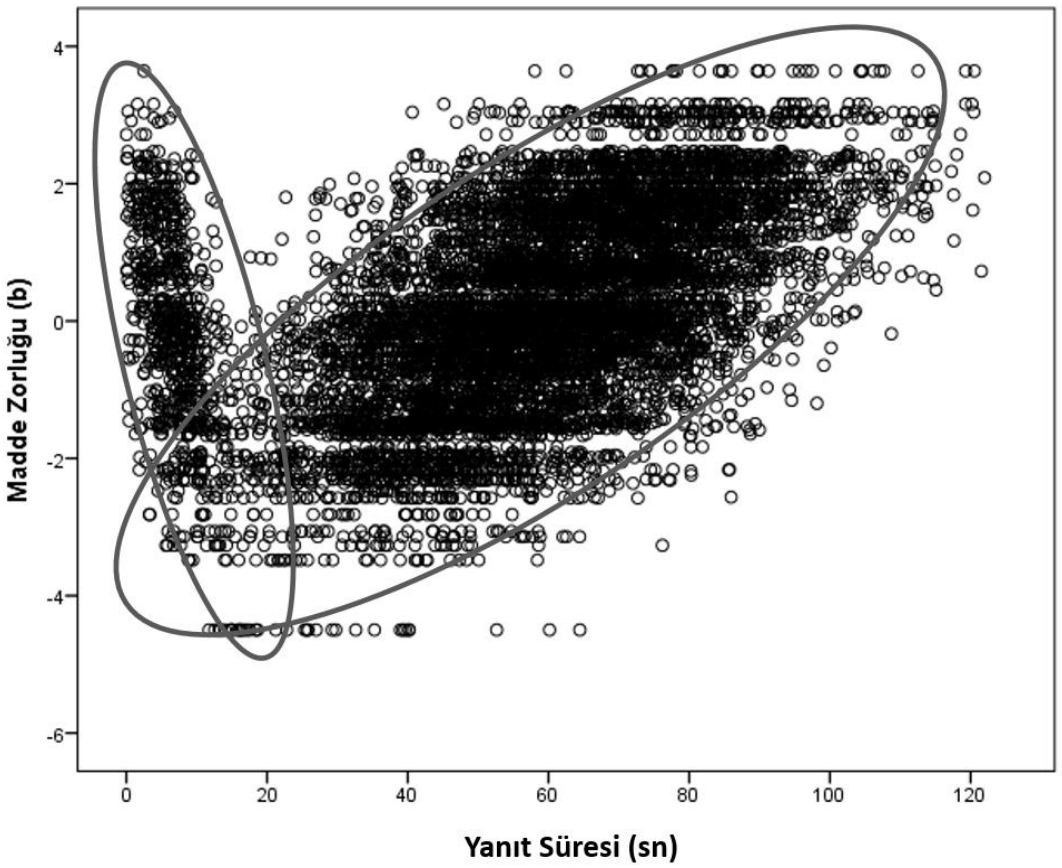
Bu bölümde, yürütülen benzetim çalışmasına ait çıktılar yorumlanmıştır. Madde kalibrasyonu, zaman türetimi, BUT 3PL, BUT ÇD ve kâğıt-kalem yöntemlerine ait uyum sonuçları sunulmuştur.

Testin geçerli ve güvenilir bir test olabilmesi için maddeler, zorluk / yetenek düzeyi ölçeğine uygun şekilde kalibre edilmelidir. Madde zorluk ve yetenek düzeyi ölçeklemesinin uygunluğunu sunmak için madde sayısı 250, standart hata 0,3 olan senaryoya ait 1000 tekrar içerisinden rastgele seçilen 50 tekrar sunulmuştur (Şekil 3.1). Benzetim çalışmasında madde zorluğu ve kişi yetenek düzeyi dağılımları uygundur.



Şekil 3.1. Rastgele seçilen 50 tekrarda madde zorluğu / yetenek düzeyi dağılımı

Literatürde belirtilen madde zorluğu ve yanıt süresi arasındaki pozitif korelasyon göz önünde bulundurularak türetilen yanıt süreleri sunulmuştur (Şekil 3.2). ÇD uygulayan kişilerde soru zorluğu ile yanıt süresi arasında pozitif ilişki (r=0,6), hızlı yanıt davranışı sergileyen kişilerde ise soru zorluğu ile yanıt süresi arasında negatif yapıda ilişki (r=-0,6) oluşturulmuştur.



Şekil 3.2. Madde zorluğu ve yanıt süresi arasındaki ilişki

BUT 3PL ve BUT ÇD uygulamalarındaki tüm benzetim senaryolarında 1000 tekrarın her birinde kullanılan ortalama madde sayıları elde edilmiş ve bu madde sayılarının senaryo bazlı ortalamaları sunulmuştur (Çizelge 3.1).

Benzetim çalışması sonucunda yöntemden bağımsız olarak soru bankasındaki madde sayısı artarken her iki standart hata düzeyi için de uygulanan ortalama soru sayısı azalmıştır. Standart hatanın 0,3'ten 0,5'e çıkması ise beklendiği gibi uygulanan

soru sayısının artmasına sebep olmuştur. Standart hataya bağlı olarak kullanılan soru sayısındaki bu artış ve madde sayısı arttıkça kullanılan soru sayısındaki azalış BUT 3PL ve BUT ÇD yöntemlerinde benzerlik göstermiştir. BUT ÇD'de uygulanan ortalama soru sayılarının BUT 3PL'ye göre daha düşük olduğu görülmüştür.

Çizelge 3.1. BUT 3PL ve BUT ÇD uygulamalarında kullanılan madde sayıları

Madde Sayısı	S. Hata	Ortalama Madde Sayısı BUT 3PL	Ortalama Madde Sayısı BUT ÇD
50	0,3	42,8±4,8 (27,1-50,0)	39,4±4,3 (25,2-46,6)
	0,5	10,8±1,9 (4,7-18,0)	10,5±1,7 (5,0-17,3)
100	0,3	38,4±5,4 (20,4-67,0)	35,5±4,8 (19,1-61,4)
	0,5	8,6±1,5 (4,8-18,5)	8,6±1,3 (5,3-17,2)
250	0,3	31,5±4,0 (19,4-43,7)	29,6±3,4 (18,5-39,1)
	0,5	7,2±0,9 (4,3-13,2)	7,5±0,8 (4,7-12,0)

Sonuçlar ortalama±standart sapma (minimum-maksimum) değer olarak sunulmuştur

BUT 3PL ve BUT ÇD uygulamalarındaki tüm benzetim senaryolarında 1000 tekrarın her birinde ortalama kestirim yanlılığı elde edilmiş ve bu yanlılık senaryo bazlı olarak sunulmuştur (Çizelge 3.2).

BUT 3PL yönteminde tüm madde sayıları için standart hata azaldıkça ortalama yanlılığın da azaldığı görüldü. Madde sayısındaki artış ortalama yanlılık üzerinde ciddi etkiye neden olmadığı halde düşük standart hatada (0,3) yanlılığın en düşük olduğu koşul 250 madde olan senaryoydu.

Çizelge 3.2. BUT 3PL ve BUT ÇD uygulamaları ile kestirilen yetenek düzeylerinin gerçek yetenek düzeyine göre yanlılığı

Madde Sayısı	S. Hata	Ortalama Yanlılık BUT 3PL	Ortalama Yanlılık BUT ÇD
50	0,3	0,017±0,053 (-0,157-0,202)	-0,296±0,074 (-0,596-0,053)
	0,5	0,117±0,060 (-0,120-0,330)	-0,168±0,073 (-0,370-0,080)
100	0,3	0,031±0,060 (-0,197-0,246)	-0,253±0,073 (-0,482-0,010)
	0,5	0,120±0,064 (-0,191-0,335)	-0,149±0,072 (-0,329-0,140)
250	0,3	0,015±0,107 (-0,362-0,295)	-0,244±0,114 (-0,639-0,036)
	0,5	0,089±0,102 (-0,325-0,392)	-0,172±0,106 (-0,531-0,192)

Sonuçlar ortalama±standart sapma (minimum-maksimum) değer olarak sunulmuştur

Çözümsel davranış ve 3PL modele ait BUT uygulamalarında yetenek düzeyi kestirimlerine ait ortalama standart hatalar elde edilmiş ve senaryo bazlı olarak sunulmuştur (Çizelge 3.3).

Ortalama standart hata için elde edilen bulgular incelendiğinde madde sayısındaki artışın her iki yöntemde de ortalama standart hatayı azalttığı sonucuna varıldı. Benzer şekilde madde sayısındaki artışın ortalama standart hatayı düşürdüğü görüldü. Senaryolardan 50 madde içeren ve standart hata 0,3 olan durumda her iki yöntemde de ortalama standart hatanın minimum değeri beklenen standart hata düzeyine ulaşamadı.

Çizelge 3.3. BUT 3PL ve BUT ÇD uygulamalarında yetenek düzeyi kestirimlerine ait standart hatalar

Madde Sayısı	S. Hata	Ortalama Standart Hata BUT 3PL	Ortalama Standart Hata BUT ÇD
50	0,3	0,411±0,032 (0,337-0,673)	0,380±0,028 (0,314-0,657)
	0,5	0,503±0,024 (0,400-0,650)	0,476±0,023 (0,380-0,600)
100	0,3	0,324±0,010 (0,287-0,377)	0,305±0,011 (0,269-0,470)
	0,5	0,471±0,012 (0,401-0,504)	0,450±0,012 (0,375-0,481)
250	0,3	0,299±0,004 (0,266-0,310)	0,286±0,005 (0,250-0,295)
	0,5	0,466±0,010 (0,396-0,480)	0,448±0,010 (0,377-0,467)

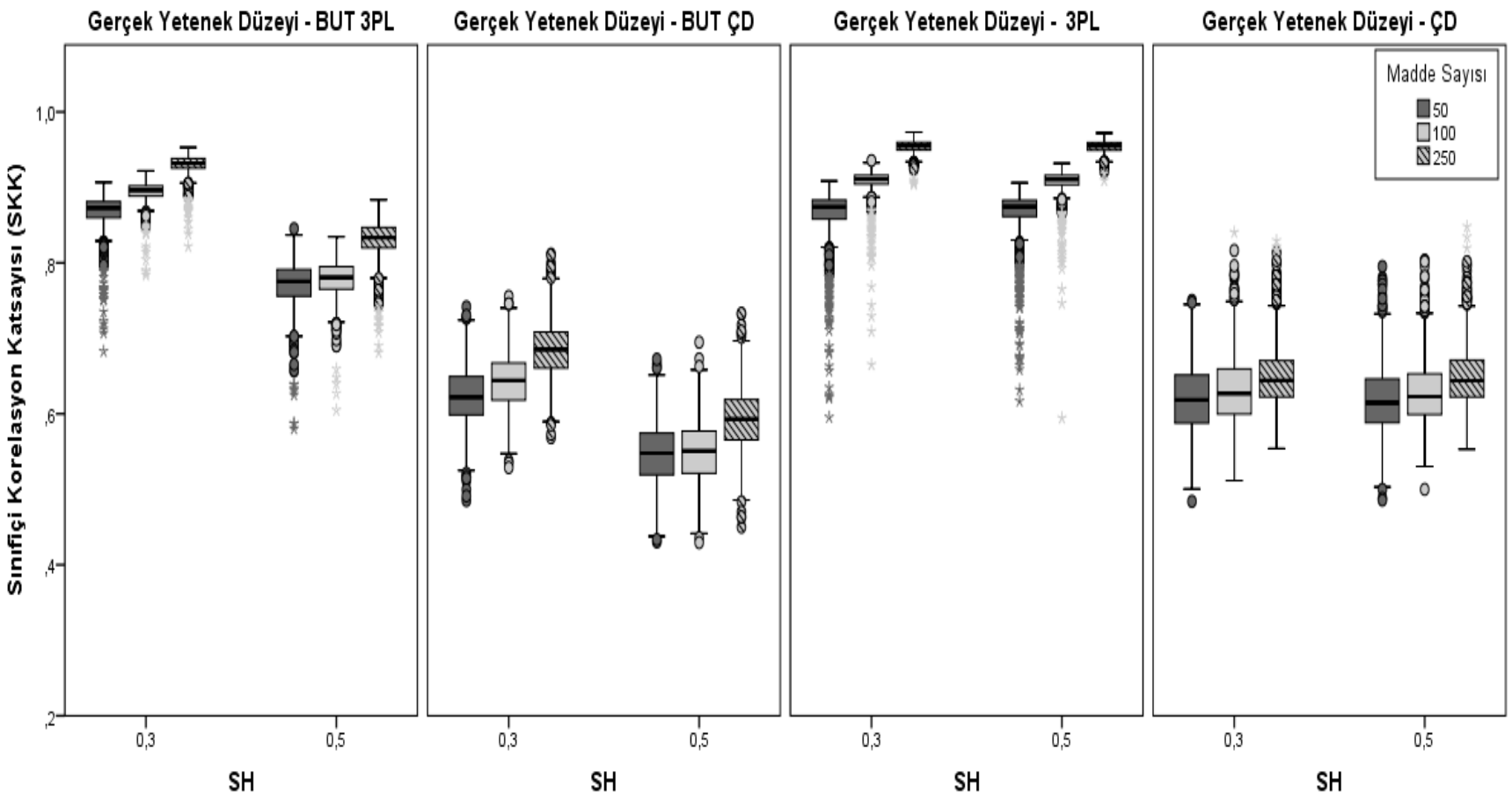
Sonuçlar ortalama±standart sapma (minimum-maksimum) değer olarak sunulmuştur

Gerçek yetenek düzeyleri ile BUT 3PL, BUT ÇD, kâğıt-kalem 3PL ve kâğıt-kalem ÇD yöntemleri kullanılarak kestirilen yetenek düzeyleri arasındaki ilişkiyi değerlendirmek için elde edilen SKK değerleri Çizelge 3.4 ve Şekil 3.3'te sunuldu. Madde sayısındaki artışın gerçek yetenek düzeyi ve BUT yöntemleri ile kestirilen yetenek düzeyleri arasındaki SKK'nın artmasına sebep olduğu görüldü. Benzer şekilde standart hatadaki azalışın SKK'nın artmasına sebep olduğu sonucuna varıldı. Kağıt-kalem yöntemiyle yapılan kestirimlerde de madde sayısındaki artış SKK'yı arttırdı. BUT 3PL ve BUT ÇD yöntemlerine ait SKK'lar incelendiğinde BUT 3PL yöntemiyle elde edilen yetenek düzeyleri ile gerçek yetemek düzeyleri arasındaki uyum iyi düzeydeyken, BUT ÇD yöntemiyle elde edilen yetenek düzeyleriyle gerçek yetenek düzeyleri arasındaki uyum orta düzeydeydi. Benzer şekilde kağıt-kalem yöntemlerinde gerçek yetenek düzeyleriyle uyumda ÇD modeliyle elde edilen yetenek düzeyleri 3PL modelle elde edilen yetenek düzeylerinden daha düşük uyuma sahipti.

Çizelge 3.4. Tüm yöntemlerle kestirilen yetenek düzeyi ile gerçek yetenek düzeyleri arasındaki SKK

Madde Sayısı	S. Hata	Gerçek Teta ve BUT 3PL	Gerçek Teta ve BUT ÇD	Gerçek Teta ve 3PL	Gerçek Teta ve ÇD
50	0,3	0,866±0,026 (0,683-0,907)	0,623±0,038 (0,485-0,742)	0,862±0,038 (0,595-0,908)	0,621±0,045 (0,484-0,750)
	0,5	0,772±0,030 (0,580-0,845)	0,547±0,040 (0,430-0,672)	0,864±0,037 (0,617-0,906)	0,619±0,046 (0,486-0,795)
100	0,3	0,894±0,014 (0,784-0,922)	0,644±0,036 (0,529-0,756)	0,907±0,020 (0,665-0,936)	0,633±0,048 (0,512-0,840)
	0,5	0,778±0,025 (0,605-0,835)	0,549±0,041 (0,430-0,695)	0,907±0,020 (0,594-0,932)	0,630±0,046 (0,500-0,804)
250	0,3	0,930±0,013 (0,822-0,953)	0,685±0,038 (0,568-0,812)	0,954±0,008 (0,904-0,973)	0,651±0,042 (0,554-0,829)
	0,5	0,831±0,025 (0,681-0,884)	0,593±0,040 (0,450-0,733)	0,954±0,008 (0,909-0,972)	0,649±0,041 (0,553-0,848)

Sınıfiçi korelasyon katsayısı (SKK) ile ilgili sonuçlar ortalama±standart sapma (minimum-maksimum) değer olarak sunulmuştur. BUT: Bilgisayar Uyarlamalı Test, 3PL: Üç Parametreli Lojistik model, ÇD: Çözümsel Davranış modeli



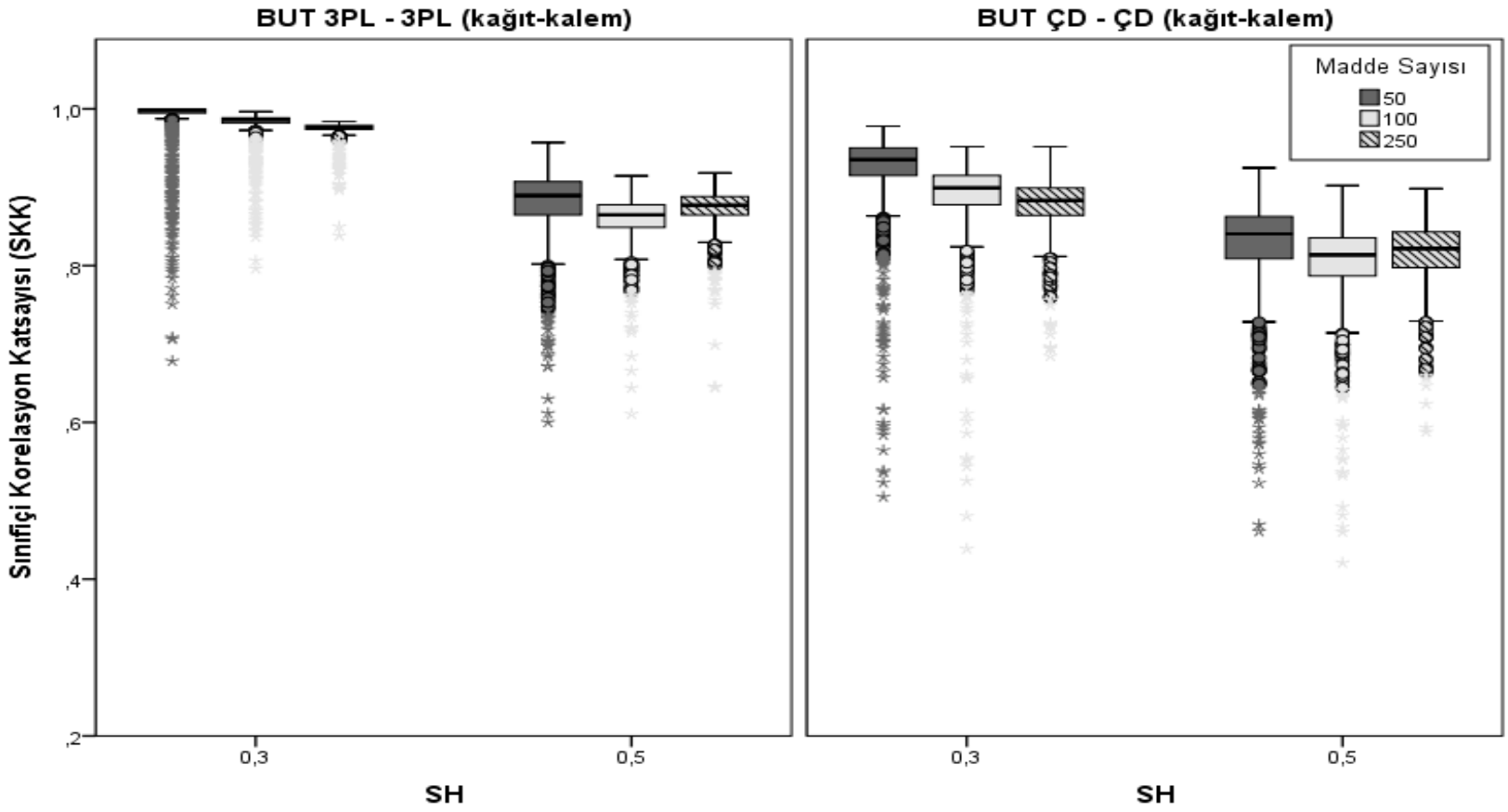
Şekil 3.3. Gerçek yetenek düzeyi ile tüm yöntemler arasındaki SKK

Kağıt-kalem yöntemi ve BUT ile kestirilen yetenek düzeylerinin ilişkisi incelendi ve elde edilen SKK'lar Çizelge 3.5 ve Şekil 3.4'te sunuldu. Tüm madde sayıları için standart hatanın azalması ile BUT 3PL ve kağıt-kalem (3PL) yöntemleri arasındaki SKK'nın arttığı görüldü. Benzer şekilde BUT ÇD ile kağıt-kalem yöntemi (ÇD) arasındaki SKK da artmaktaydı. Kağıt-kalem yöntemleri ve BUT yöntemleri ile kestirilen yetenek düzeyleri arasındaki uyum mükemmel düzeydedir.

Çizelge 3.5. Kağıt-kalem yöntemi ve BUT ile kestirilen yetenek düzeyleri arasındaki SKK

Madde Sayısı	S. Hata	BUT 3PL ve 3PL	BUT ÇD ve ÇD
50	0,3	0,984±0,040 (0,678-1,0)	0,919±0,060 (0,505-0,978)
	0,5	0,878±0,047 (0,600-0,957)	0,827±0,060 (0,461-0,925)
100	0,3	0,980±0,022 (0,796-0,996)	0,888±0,051 (0,439-0,952)
	0,5	0,859±0,030 (0,611-0,914)	0,804±0,052 (0,421-0,902)
250	0,3	0,974±0,012 (0,838-0,984)	0,879±0,033 (0,685-0,952)
	0,5	0,873±0,024 (0,644-0,918)	0,817±0,039 (0,588-0,898)

Sınıfiçi korelasyon katsayısı (SKK) ile ilgili sonuçlar ortalama±standart sapma (minimum-maksimum) değer olarak sunulmuştur. BUT: Bilgisayar Uyarlamalı Test, 3PL: Üç Parametreli Lojistik model, ÇD: Çözümsel Davranış modeli



Şekil 3.4. BUT ve kağıt-kalem yöntemi ile kestirilen yetenek düzeyleri arasındaki SKK

3.1. BUT 3PL ve BUT ÇD Kestirimlerinin Gerçek Yetenek Düzeyi ile Uyumlarının Karşılaştırılması

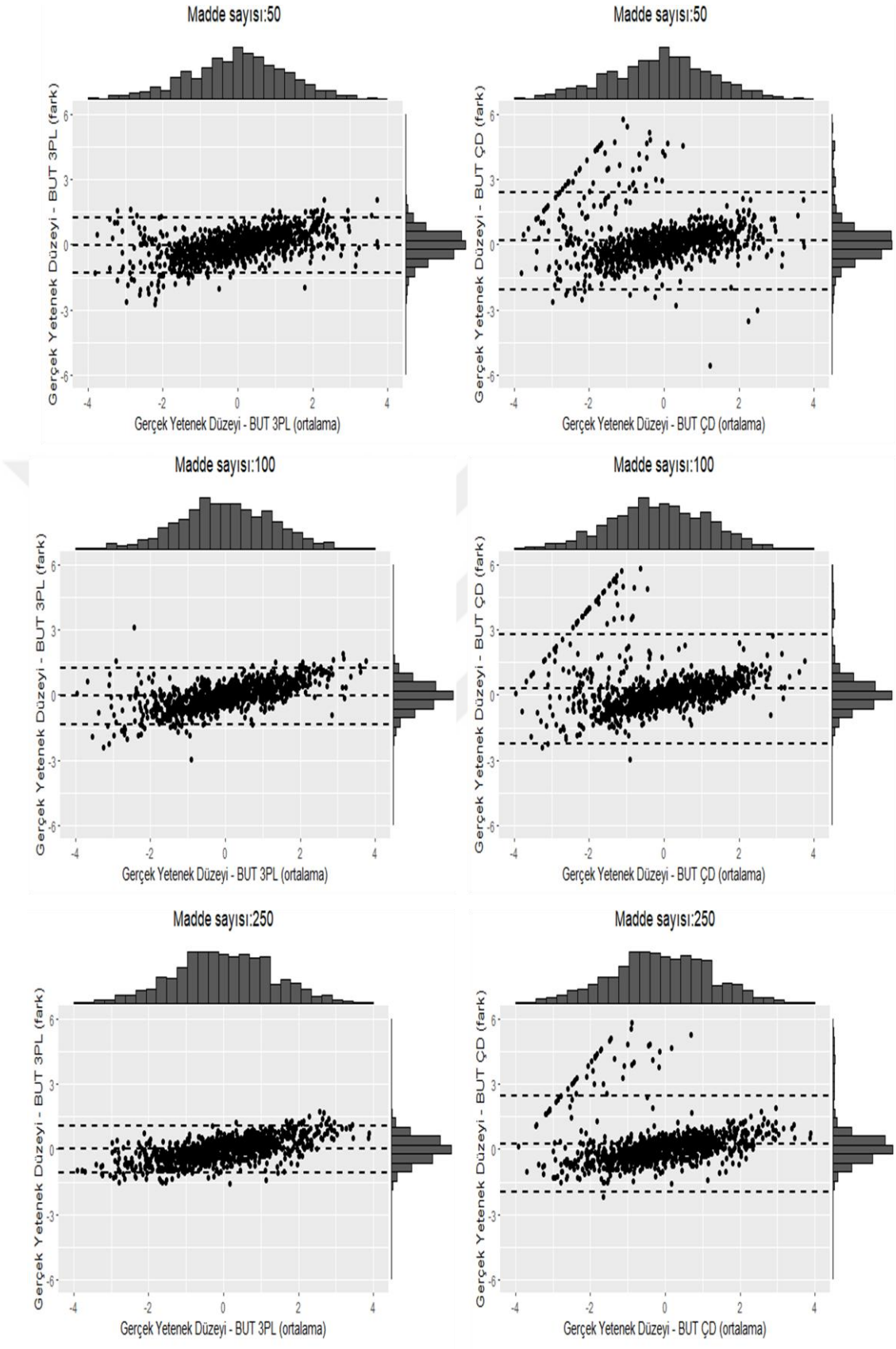
Standart hata ve madde sayısına bağlı olarak BUT yöntemleri ile gerçek yetenek düzeyleri arasındaki uyum incelendi ve uyum sınırları dışında kalan kişi yüzdeleri sunuldu.

Standart hatanın 0,3 olduğu durumda, BUT 3PL yönteminde madde sayısı arttıkça uyum sınırları daralmış ve uyum sınırları dışında kalan kişilerin yüzdesi azalmıştır. BUT 3PL'ye benzer şekilde BUT ÇD yönteminde de durum benzerdi (Şekil 3.5, Çizelge 3.6).

Çizelge 3.6. BUT yöntemleriyle kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları (SH:0,3)

Madde Sayısı	Gerçek yetenek düzeyi ve BUT 3PL	Uyum sınırlarının dışında kalan kişi yüzdesi	Gerçek yetenek düzeyi ve BUT ÇD	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,017 (-1,512; 1,545)	%5,1	-0,296 (-2,883;2,313)	%6,1
100	0,031 (-1,332;1,391)	%3,9	-0,253 (-2,717;2,229)	%5,8
250	0,015 (-1,095;1,123)	%3,8	-0,244 (-2,563;2,084)	%5,4

Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.



Şekil 3.5. Madde sayısı değiştiğinde BUT 3PL ve BUT ÇD'nin gerçek yetenek düzeyi ile uyumlarının karşılaştırılması (S.H:0,3)

Gerçek yetenek düzeyi ve BUT uygulamaları arasındaki fark belirlenen sınırlara göre (0,3 birim, 0,5 birim ve 1 birim) değerlendirildi ve sınırlar dışında kalan kişi yüzdeleri sunuldu. Hem BUT 3PL hem de BUT ÇD yöntemlerinde madde sayısı arttıkça belirlenen tüm sınırlar dışında kalan kişi yüzdelerinin azaldığı görüldü. Madde sayısının 250 olduğu senaryoda BUT 3PL yöntemi ile yapılan kestirimlerde gerçek yetenek düzeyinden 1 birimlik farklılaşma %5 seviyesine indiği halde, BUT ÇD yönteminde 1 birimlik sınır dışında kalan kişi yüzdesi %12,2 idi (Çizelge 3.7).

Çizelge 3.7. BUT yöntemleriyle kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı (SH:0,3)

Madde Sayısı	**Gerçek yetenek düzeyi ve BUT 3PL Belirlenen birimler dışında kalan kişi yüzdesi**			**Gerçek yetenek düzeyi ve BUT ÇD Belirlenen birimler dışında kalan kişi yüzdesi**		
	0,3	0,5	1,0	0,3	0,5	1,0
50	%62,5	%42,3	%13,6	%66,3	%48,0	%21,4
100	%59,8	%38,6	%10,3	%63,4	%43,8	%17,4
250	%52,8	%30,0	%5,6	%56,6	%35,5	%12,2

Standart hatanın 0,5 olduğu durumda, BUT 3PL yönteminde madde sayısı arttıkça uyum sınırları daralmış fakat uyum sınırları dışında kalan kişilerin yüzdesinde ciddi bir değişiklik olmamıştır. BUT ÇD yönteminde de durum BUT 3PL ile benzerdi. Tüm madde sayılarında BUT 3PL yönteminde uyum sınırları BUT ÇD yöntemine göre daha dar ve sınırlar dışında kalan kişi yüzdeleri daha düşüktü (Şekil 3.6, Çizelge 3.8).

Çizelge 3.8. BUT yöntemleriyle kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları (SH:0,5)

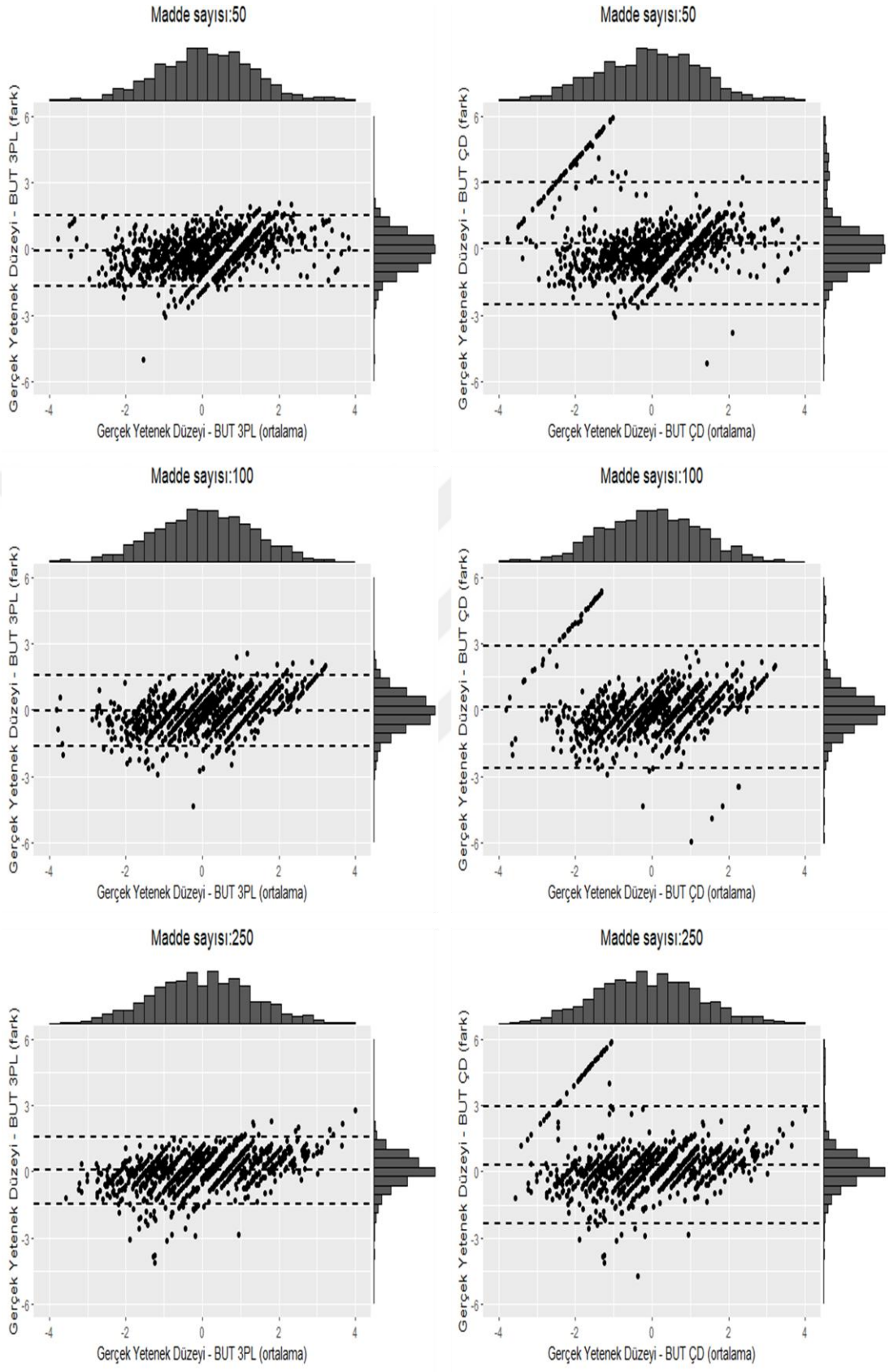
Madde Sayısı	Gerçek yetenek düzeyi ve BUT 3PL (ortalama fark ve %95 G.A.)	Uyum sınırlarının dışında kalan kişi yüzdesi	Gerçek yetenek düzeyi ve BUT ÇD (ortalama fark ve %95 G.A.)	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,117 (-1,695;1,923)	%5,0	-0,168 (-2,949; 2,621)	%6,4
100	0,120 (-1,657;1,890)	%5,0	-0,149 (-2,891; 2,602)	%6,4
250	0,089 (-1,535;1,708)	%4,9	-0,172 (-2,832; 2,498)	%6,3

Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.

Standart hata 0,5 olan durumda da 0,3 olan duruma benzer şekilde gerçek yetenek düzeyi ve BUT uygulamaları arasındaki fark belirlenen sınırlara göre değerlendirildi ve sınırlar dışında kalan kişi yüzdeleri sunuldu. Hem BUT 3PL hem de BUT ÇD yöntemlerinde madde sayısı arttıkça belirlenen tüm sınırlar dışında kalan kişi yüzdelerinin azaldığı görüldü. Fakat her koşulda bu kişi yüzdeleri standart hata 0,3 olan durumdaki yüzdelerden daha yüksekti. (Çizelge 3.9) .

Çizelge 3.9. BUT yöntemleriyle kestirilen yetenek düzeyleri ile gerçek yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı (SH:0,5)

Madde Sayısı	Gerçek yetenek düzeyi ve BUT 3PL Belirlenen birimler dışında kalan kişi yüzdesi			Gerçek yetenek düzeyi ve BUT ÇD Belirlenen birimler dışında kalan kişi yüzdesi		
	0,3	0,5	1,0	0,3	0,5	1,0
50	%70,2	%52,7	%22,6	%72,5	%56,3	%28,2
100	%69,1	%51,1	%21,2	%71,3	%54,7	%26,6
250	%65,1	%45,7	%16,5	%67,6	%49,6	%22,1



Şekil 3.6. Madde sayısı değiştiğinde BUT 3PL ve BUT ÇD'nin gerçek yetenek düzeyi ile uyumlarının karşılaştırılması (S.H:0,5)

3.2. BUT 3PL ve BUT ÇD Uygulamaları ile Kestirilen Yetenek Düzeylerinin Uyumu

Üç parametreli model ile çözümsel davranış modelinin BUT uygulamaları ile kestirilen yetenek düzeyleri arasındaki fark ve uyum sınırları dışında kalan kişi yüzdeleri hesaplandı (Şekil 3.7, Çizelge 3.10). Madde sayısı arttıkça yöntemlerle kestirilen yetenek düzeyleri arasındaki farka ait güven aralıklarının azaldığı gözlendi. Bu durum standart hatanın 0,3 ve 0,5 olduğu durumlarda benzerdi.

Çizelge 3.10. BUT 3PL ve BUT ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları

Madde Sayısı	BUT 3PL ve BUT ÇD (SH:0,3)	Uyum sınırlarının dışında kalan kişi yüzdesi	BUT 3PL ve BUT ÇD (SH:0,5)	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,301 (-1,778;2,381)	%6,6	0,278 (-1,827;2,383)	%6,4
100	0,274 (-1,742;2,289)	%6,1	0,261 (-1,809;2,331)	%6,2
250	0,253 (-1,734;2,241)	%5,4	0,254 (-1,843;2,351)	%5,9

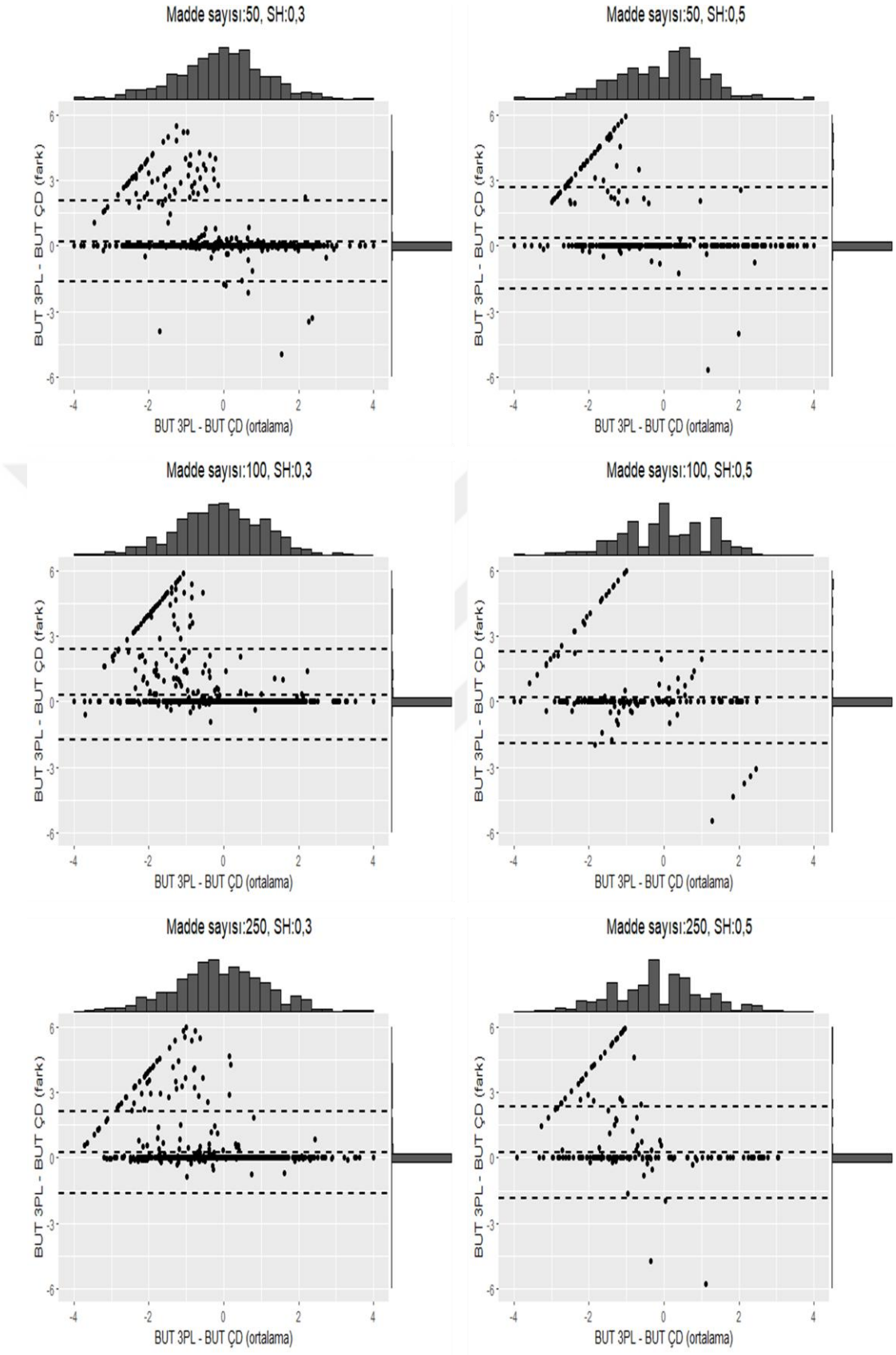
Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.

BUT 3PL ve BUT ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı incelendiğinde, madde sayısı düşükken standart hatanın 0,5 olduğu durumda yöntemler arasındaki farka ait kişi yüzdelerinin 0,3 olan durumdan daha düşük olduğu fakat madde sayısı arttıkça bu yüzdelerin benzer duruma geldiği görüldü (Çizelge 3.11). En düşük madde sayısı içeren senaryoda bile yöntemlerle kestirilen yetenek düzeyleri arasındaki farkın belirlenen en düşük sınırın (0,3 birim) içinde olma yüzdesi %88,3 idi.

Çizelge 3.11. BUT 3PL ve BUT ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı

Madde Sayısı	BUT 3PL ve BUT ÇD (SH:0,3) Belirlenen birimler dışında kalan kişi yüzdesi			BUT 3PL ve BUT ÇD (SH:0,5) Belirlenen birimler dışında kalan kişi yüzdesi		
	0,3	0,5	1,0	0,3	0,5	1,0
50	%11,7	%10,7	%9,4	%8,9	%8,4	%7,7
100	%10,7	%9,7	%8,2	%8,5	%7,9	%7,2
250	%9,6	%8,6	%7,2	%8,4	%7,8	%7,1

Bazı kestirimlerin uyum sınırları dışında kaldığı gözlendiği halde Y-eksenine ait histogram incelendiğinde bu farklılıkların çok düşük frekansa sahip olduğu ve genel olarak yöntem kestirim farklarının sıfır etrafında dağıldığı görüldü (Şekil 3.7).



Şekil 3.7. Madde sayısı ve standart hata değiştiğinde BUT 3PL ve BUT ÇD'nin yetenek düzeyi kestirimlerinde uyumu

3.3. BUT 3PL ve 3PL Model ile Kestirilen Yetenek Düzeylerinin Uyumu

Üç parametreli model BUT uygulaması ve 3PL kâğıt-kalem yöntemine ait kestirimlerin uyumu incelendi ve Şekil 3.8'de sunuldu. Standart hata 0,5 iken hatanın 0,3 olduğu duruma göre ortalama fark daha yüksek, uyum sınırları daha geniş ve sınırlar içerisinde kalan kişi yüzdesi daha azdır. Madde sayısındaki artış uyum sınırları dışında kalan kişi yüzdesini arttırmıştır. Bunun sebebi soru bankasında 50 madde sayısına ve 0,3 standart hata durdurma kriterine sahip BUT senaryosunda sürecin sonlanmaması ve neredeyse tüm sorularında kişiye yönlendirilmesidir. Bu sebeple düşük madde sayısına sahip senaryoda standart hata da düşükken sınırlar dışında kalan kişi yüzdesi (%1) en düşüktü (Çizelge 3.12).

Çizelge 3.12. BUT 3PL ve 3PL yöntemleriyle kestirilen yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları

Madde Sayısı	BUT 3PL ve 3PL (SH:0,3)	Uyum sınırlarının dışında kalan kişi yüzdesi	BUT 3PL ve 3PL (SH:0,5)	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,011 (-0,418;0,439)	%1	0,107 (-0,982;1,197)	%4,2
100	0,022 (-0,417;0,461)	%3,3	0,109 (-1,007;1,225)	%4,9
250	0,029 (-0,498;0,558)	%4,6	0,102 (-1,047;1,251)	%5,0

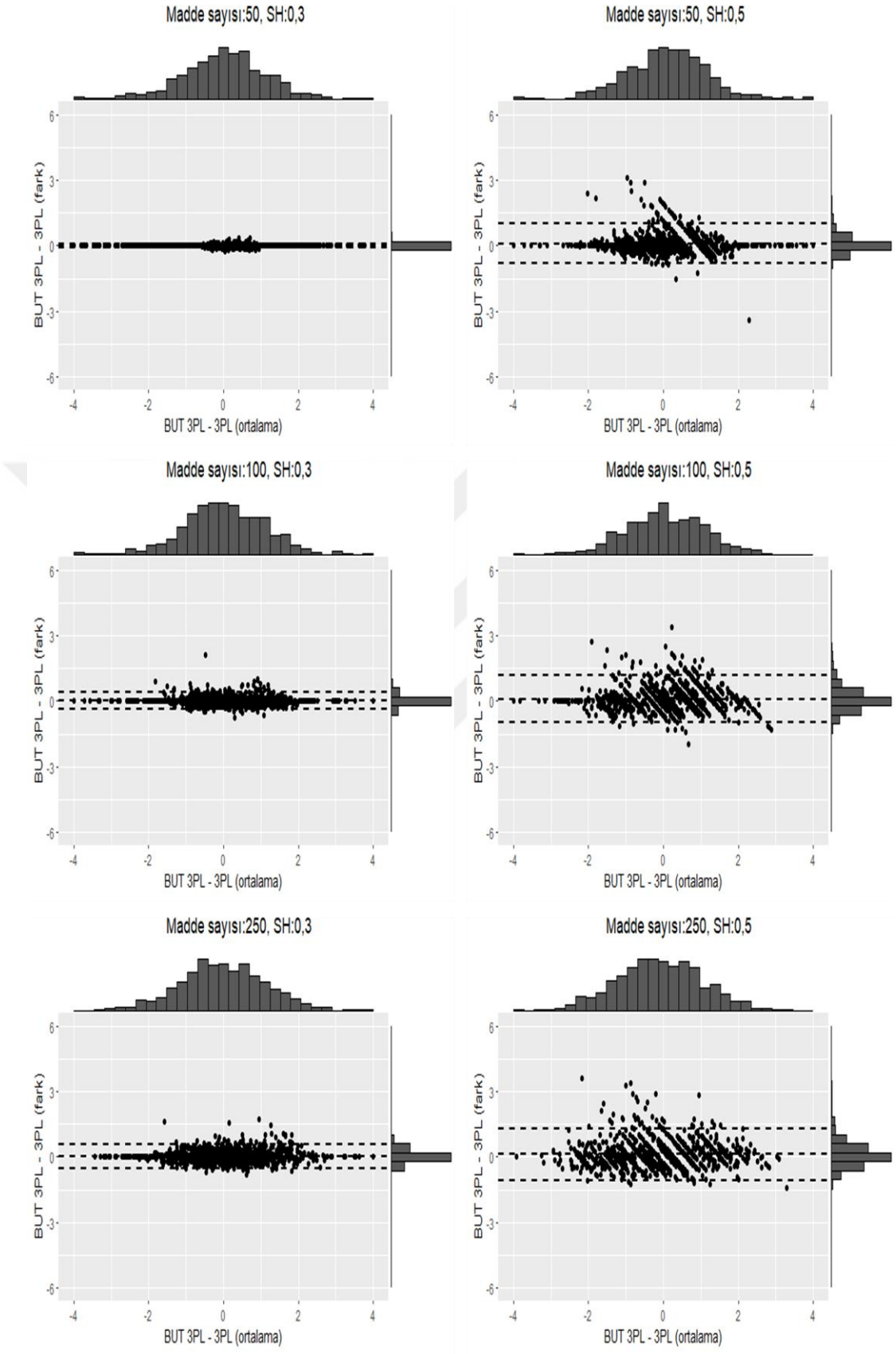
Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.

Benzer şekilde belirlenen sınırlara göre değerlendirme yapıldığında madde sayısı artarken uyum sınırları dışında kalan kişi yüzdesi artmaktadır. Her koşulda standart hatanın düşük olması uyum sınırları içerisinde kalan kişi yüzdesini standart hata yüksek olan duruma göre arttırmıştır (Çizelge 3.13).

Çizelge 3.13. BUT 3PL ve 3PL yöntemleriyle kestirilen yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı

Madde Sayısı	BUT 3PL ve 3PL (SH:0,3) Belirlenen birimler dışında kalan kişi yüzdesi			BUT 3PL ve 3PL (SH:0,5) Belirlenen birimler dışında kalan kişi yüzdesi		
	0,3	0,5	1,0	0,3	0,5	1,0
50	%1,6	%0,9	%0,3	%37,2	%19,5	%5,6
100	%9,3	%2,4	%0,4	%46,9	%26,2	%6,9
250	%20,8	%5,7	%0,5	%52,2	%30,4	%7,7

Grafik incelendiğinde standart hatanın 0,5 olmasının saçılımı arttırdığı ve güven aralığını genişlettiği görüldü. Düşük standart hatada ise genel olarak BUT ve kağıt-kalem kestirim farklarının sıfır etrafında dağıldığı görüldü (Şekil 3.8).



Şekil 3.8. Madde sayısı ve standart hata değiştiğinde BUT 3PL ve 3PL (kağıt-kalem yöntemi)'nin yetenek düzeyi kestirimlerinde uyumu

3.4. BUT ÇD ve ÇD Model ile Kestirilen Yetenek Düzeylerinin Uyumu

BUT ÇD uygulaması ve ÇD kâğıt-kalem yöntemine ait kestirimlerin uyumu incelendi ve Şekil 3.9'da sunuldu. Üç parametreli klasik modele benzer şekilde ÇD için de standart hata 0,5 iken hatanın 0,3 olduğu duruma göre ortalama fark daha yüksek, uyum sınırları daha geniş ve sınırlar içerisinde kalan kişi yüzdesi daha azdı. Madde sayısındaki artış uyum sınırları dışında kalan kişi yüzdesini arttırdı (Çizelge 3.14).

Çizelge 3.14. BUT ÇD ve ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları

Madde Sayısı	BUT ÇD ve ÇD (SH:0,3)	Uyum sınırlarının dışında kalan kişi yüzdesi	BUT ÇD ve ÇD (SH:0,5)	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,032 (-1,155;1,219)	%4,4	0,159 (-1,523;1,841)	%5,6
100	0,103 (-1,237;1,442)	%5,0	0,210 (-1,533;1,953)	%5,5
250	0,167 (-1,276;1,610)	%5,3	0,235 (-1,531;2,001)	%5,3

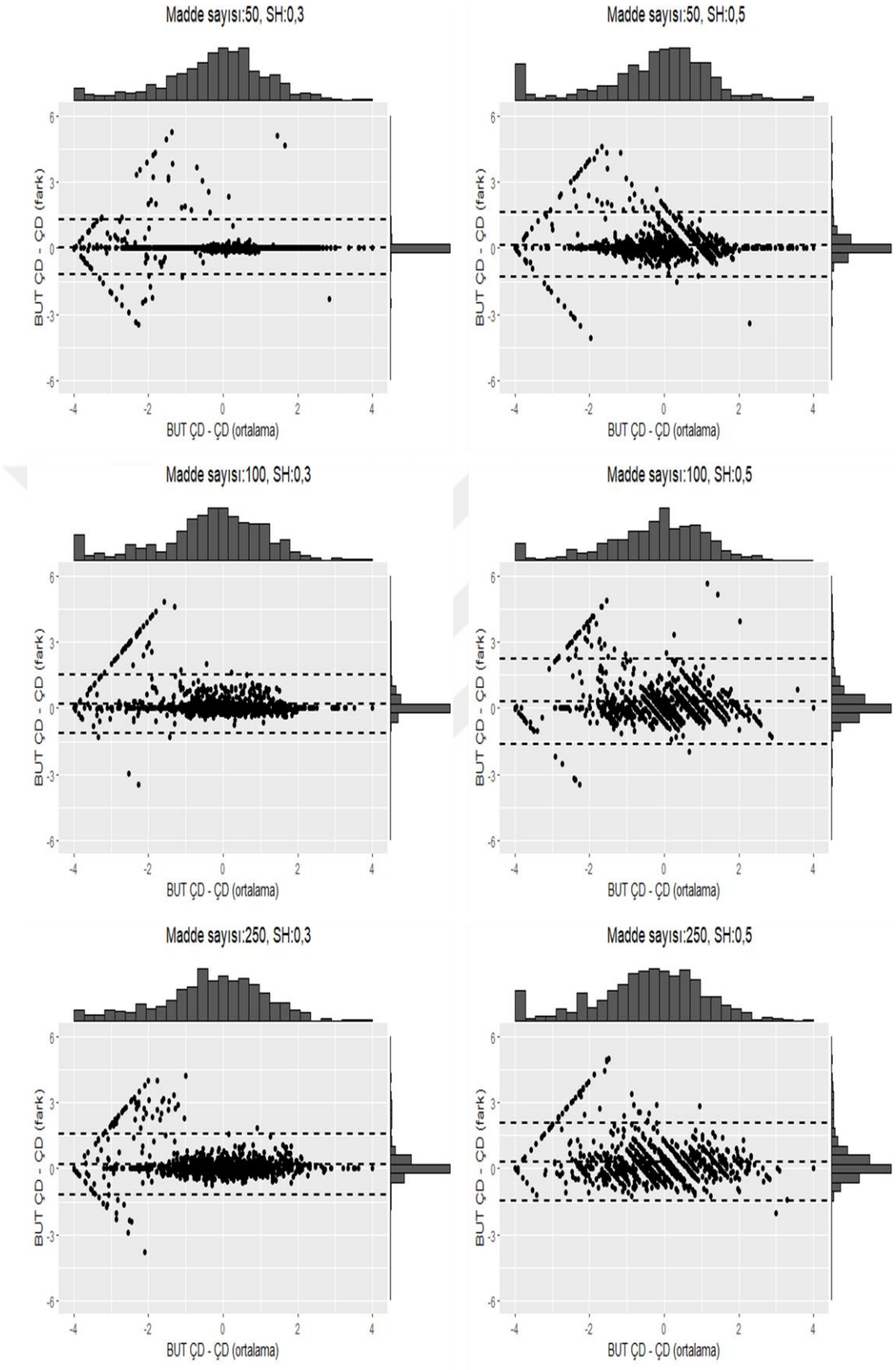
Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.

Belirlenen sınırlara göre değerlendirme yapıldığında üç parametreli klasik modele benzer şekilde ÇD modelinde de madde sayısı artarken uyum sınırları dışında kalan kişi yüzdesi arttı. Her koşulda standart hatanın düşük olması uyum sınırları içerisinde kalan kişi yüzdesini standart hata yüksek olan duruma göre arttırmıştır (Çizelge 3.15).

Çizelge 3.15. BUT ÇD ve ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı

Madde Sayısı	BUT ÇD ve ÇD (SH:0,3) Belirlenen birimler dışında kalan kişi yüzdesi			BUT ÇD ve ÇD (SH:0,5) Belirlenen birimler dışında kalan kişi yüzdesi		
	0,3	0,5	1,0	0,3	0,5	1,0
50	%8,9	%7,2	%4,9	%41,3	%24,9	%10,9
100	%18,5	%10,9	%6,4	%50,8	%31,9	%12,8
250	%29,5	%14,8	%7,1	%54,9	%35,2	%13,2

Grafik incelendiğinde standart hatanın yüksek olmasının saçılımı arttırdığı ve güven aralığını genişlettiği görüldü. Düşük standart hatada ise genel olarak BUT ve kağıt-kalem kestirim farklarının sıfır etrafında dağıldığı görüldü, X eksenine ait histogram incelendiğinde en düşük yetenek düzeyinde (teta:-4) bir yığılım olduğu görüldü (Şekil 3.9).



Şekil 3.9. Madde sayısı ve standart hata değiştiğinde BUT ÇD ve ÇD (kağıt-kalem yöntemi)'nin yetenek düzeyi kestirimlerinde uyumu

3.5. 3PL ve ÇD ile Kestirilen (Kâğıt-Kalem Yöntemleri) Yetenek Düzeylerinin Uyumu

Kağıt-kalem yöntemleri arasındaki uyum incelendiğinde madde sayısı arttıkça uyum sınırları içinde kalan kişi yüzdesinin azaldığı ve yöntemlerle kestirilen yetenek düzeyleri arasındaki farka ait güven aralıklarının arttığı görüldü (Şekil 3.10, Çizelge 3.16).

Çizelge 3.16. 3PL ve ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki ortalama fark ve Bland-Altman uyum sınırları

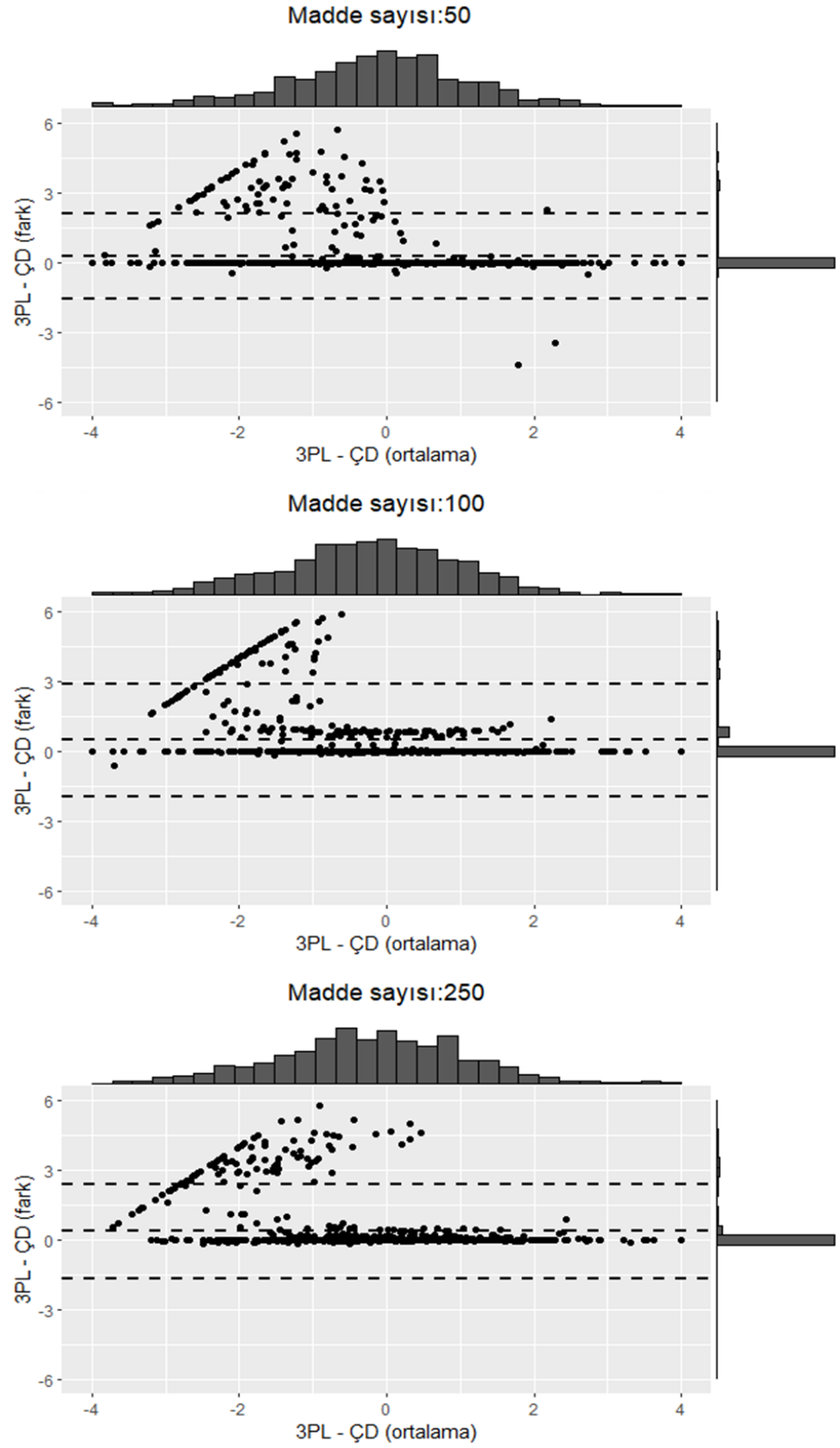
Madde Sayısı	3PL ve ÇD	Uyum sınırlarının dışında kalan kişi yüzdesi
50	0,323 (-1,792;2,438)	%7,5
100	0,354 (-1,824;2,533)	%7,7
250	0,391 (-1,898;2,679)	%8,3

Sonuçlar ortalama fark ve %95 G.A. olarak sunulmuştur.

Fakat belirlenen sınırlarda kişi yüzdeleri incelendiğinde madde sayısındaki değişimin uyum sınırları içinde kalan kişi yüzdesi üzerinde ciddi bir etkiye sahip olmadığı görüldü (Çizelge 3.17).

Çizelge 3.17. 3PLve ÇD yöntemleriyle kestirilen yetenek düzeyleri arasındaki farkın belirlenen sınırlara göre dağılımı

Madde Sayısı	3PL ve ÇD Belirlenen birimler dışında kalan kişi yüzdesi		
	0,3	0,5	1,0
50	%12,7	%11,8	%10,7
100	%14,2	%12,9	%10,8
250	%14,5	%12,7	%10,3



Şekil 3.10. Madde sayısı ve standart hata değiştiğinde 3PL ve ÇD (kağıt-kalem yöntemi)'nin yetenek düzeyi kestirimlerinde uyumu

4. TARTIŞMA

Teknoloji ve bilgi çağındaki gelişmelere paralel olarak kâğıt-kalem testleri yerini BUT'a bırakmaktadır. Gelişen bilgisayar teknolojisi yardımıyla bu testlerde yalnızca kişinin yanıt deseni ile ilgili bilgi elde edilmez, ayrıca yanıt süreleri ve kişinin maddeyi yanıtladığı andaki davranış şekli ile ilgili bilgi toplanabilir (Lee ve Chen, 2011). Literatürde YS'lerin; test tasarımını, BUT'ta madde seçimini etkileyebileceği ve anormal yanıt tespitini iyileştirmeye yardımcı olabileceği gösterilmiştir (Wise ve Kong, 2005; Wise ve DeMars, 2006). Bu tez çalışmasında yanıt süresini içeren modelin ve klasik 3PL modelin BUT performansları karşılaştırılmış ve ana başlıklar altında tartışılmıştır. Bununla birlikte testlerin kâğıt-kalem yöntemleri ile uyumları da sunulmuştur.

4.1. Madde Sayısı ve Standart Hata

BUT'ta kullanılan madde sayısı ve bu maddeleri yanıtlamak için gereken süre geleneksel kâğıt-kalem testlerine göre daha azdır (Madsen, 1991; Wainer, 1990; Thompson ve Weiss, 2011). Bu çalışmada ön görüldüğü gibi BUT uygulamasında madde sayısının klasik teste göre çok daha az olduğu görülmüştür. Uygulanan madde sayısı BUT durdurma kriteri olarak seçilen standart hatanın düzeyinden etkilenmiştir. Standart hata düştükçe sorulan madde sayısı artmaktadır ve bu beklenen bir durumdur, literatürde bu durumu bildiren çalışmalar mevcuttur (Hol, 2008). YS içeren BUT ÇD ve klasik BUT 3PL yöntemlerinin ikisinde de soru bankasındaki madde sayısı arttıkça kişiye uygulanan madde sayısı azalmıştır. Bunun sebebi Cheng ve Twu'nun (1998) belirttiği gibi soru bankasında yer alan kişi yetenek düzeyine uygun zorluktaki madde sayısının artması olarak düşünülebilir. Tüm senaryolarda BUT ÇD'de uygulanan madde sayısı BUT 3PL modele göre daha azdır. Bu durum BUT ÇD modelinin BUT 3PL modeline göre bir üstünlüğüdür. Ancak BUT ÇD'deki yanlılığın BUT 3PL modele göre daha fazla olması modelin zayıf yönüdür.

Soru bankasındaki madde sayısı düşükken (50 madde) BUT uygulamalarında standart hataya (0,3) bağlı algoritma durdurma kriteri sağlanmadığından tüm sorular

kişilere büyük bir oranda uygulanmıştır. Sonuç olarak ortalama standart hata belirlenen kritere ulaşamadan süreç sonlanmıştır. Soru bankasındaki madde sayısının artmasıyla gerçek yetenek düzeyleri ve kestirilen yetenek düzeyleri arasındaki uyum düzeyinin tüm yöntemler için yükseldiği tespit edilmiştir. Madde sayısına ait üç senaryoda da (50,100 ve 250) gerçek yetenek düzeyleri ve kestirilen yetenek düzeyleri arasındaki uyum sınırları BUT 3PL yönteminde BUT ÇD yöntemine göre daha dar, sınırlar dışında kalan kişi oranı da daha düşük bulunmuştur.

4.2. BUT ve Kağıt-Kalem

BUT ve kağıt-kalem yöntemlerinin test performasını ve madde kestirimlerini karşılaştırmak için yapılan çalışmaların sonuçları incelendiğinde bazı farklılıklara rastlanmıştır. Bennet ve ark. (2008) BUT ve kâğıt-kalem testlerinde test performansı (yetenek düzeyi) açısından anlamlı farklılıklar bulurken, Threlfall ve ark. (2007) bu farkın minimal olduğunu, Wang ve ark. (2008) ise test modunun yetenek düzeyi üzerinde etkili olmadığını belirlemiştir. Bu tez çalışmasında standart hata 0,3 iken Wang ve arkadaşlarının yaptığı çalışmaya benzer şekilde kişi yetenek düzeyleri kestirimleri 3PL kâğıt-kalem modeli ile BUT 3PL arasında ve ÇD kâğıt-kalem ile BUT ÇD arasında benzerdi. Kağıt-kalem ve BUT yöntemi ile elde edilen kestirimler arasındaki farklar sıfır etrafında yığılma gösterdi. Fakat standart hata 0,5 iken Bennet ve arkadaşlarının çalışmasına benzer şekilde BUT ve kâğıt-kalem testleri arasında yetenek düzeyi kestirimleri açısından anlamlı farklılıklara rastlandı. BUT ve kağıt-kalem yöntemi ile kestirilen yetenek düzeylerinde standart hatanın azalması ile BUT 3PL ve kağıt-kalem (3PL) yöntemleri arasındaki SKK'nın artmasına neden oldu. Benzer şekilde BUT ÇD ile kağıt-kalem yöntemi (ÇD) arasındaki SKK da artmaktaydı. Kağıt-kalem yöntemleri ve BUT yöntemleri ile kestirilen yetenek düzeyleri arasındaki uyum mükemmel düzeydeydi. BUT 3PL uygulaması ve 3PL kâğıt-kalem yöntemine ait kestirimlerde standart hatanın yüksek olduğu durumda ortalama fark daha yüksek, uyum sınırları daha geniş ve sınırlar içerisinde kalan kişi yüzdesi daha azdı. Madde sayısındaki artış uyum sınırları dışında kalan kişi yüzdesini arttırmıştır. Bunun sebebi soru bankasında 50 madde sayısına ve 0,3 standart hata durdurma kriterine sahip BUT senaryosunda sürecin sonlanmaması ve

neredeyse tüm maddelerin kişiye yönlendirilmesidir. Bu durumda BUT uygulaması ve kağıt-kalem testi neredeyse aynıdır.

4.3. Madde Parametreleri ve Maddelerin Fiziksel Özellikleri

Madde kalibrasyonu bir BUT'un performansını etkileyen önemli faktördür. Maddeler, madde zorluğu / yetenek düzeyi ölçeğine uygun şekilde kalibre edilmezse test geçerli ve güvenilir olmaz. Bu benzetim çalışmasında madde zorluğu ve kişi yetenek düzeyi dağılımları uygundur.

BUT'ta süre ile ilgili parametrelerin (madde yavaşlık ve madde yanıt süresi gibi), madde zorluk parametresi ile güçlü bir korelasyona sahip olduğu Wang ve Hanson (2005) tarafından bildirilmiştir. Literatürde ayrıca yanıt süresinin madde zorluğu, kişi yeteneği ve test yapma stratejileri gibi değişkenlerle ilişkili olduğu da bildirilmiştir (Jansen, 1997; Parshall ve ark. 1994; Schnipke ve Scrams, 1997; Yang ve ark. 2002). Tüm bu araştırmalar yanıt süresini bağımlı veya sonuç değişkeni olarak ele almıştır. Yanıt doğruluğu ve yanıt süresi, genel olarak tek bir modele dahil edilmemiştir. Halkitis ve ark. (1996) madde zorluğunun ve maddenin fiziksel özelliklerinin (sorunun uzunluğu, karakter sayısı, grafiksel gösterim, gramatik yapı vb.) yanıt süresi ile ilişkili olduğu sonucuna ulaşmışlardır. Benzer şekilde yanıt gecikmesinin, madde zorluğundan ve maddenin içeriğinden eş zamanlı olarak etkilendiği bilinmektedir (Masters ve ark. 2005; Smith 2000; Bridgeman ve Cline 2000). Chae ve ark. (2019) tarafından yapılan gerçek veri seti içeren bir çalışmada ise madde ayırt ediciliği ve madde zorluğu ile yanıt süresi arasında negatif ilişki tespit edilmiştir. Ancak bu çalışmada kullanılan veri seti düşük örneklem ölçümünde olduğu için sonuçların genellenemeyeceği bildirilmiştir. Literatürde ÇD modeli ile ilgili BUT uygulaması bulunmamaktadır. Bu sebeple önceki çalışmalar göz önünde bulundurularak "madde zorluğu-yanıt süresi ilişkisi" içeren BUT ÇD ile yanıt süresi içermeyen BUT 3PL uygulamasının performasları karışılaştırılmıştır. Bu benzetim çalışmasında, BUT 3PL ve BUT ÇD yöntemlerinin performansları arasında uyum yüksek yüzeydedir. Fakat BUT ÇD ile yapılan yetenek düzeyi kestirimleri ile BUT 3PL kestirimlerine ait ortalama fark göz önünde bulundurulduğunda, ÇD modelinin

yetenek düzeyini daha düşük kestirmeye meyilli olduğu görülmüştür. Literatürde ÇD modeli ile ilgili BUT uygulaması bulunmamaktadır fakat Bridgeman ve Cline (2000) tarafından yapılan bir çalışmada yanıt süresinin daha uzun olması beklenen maddelerde kişinin genel yetenek düzeyiyle yanıt süresi arasında bir ilişki olmadığı tespit edilmiştir. Buradan yola çıkarak hızlı tahmin yönteminde süreye ait eşik değeri kullanarak yetenek düzeyini kestiren ÇD modeline ek olarak çok uzun süreli yanıtlar için de farklı bir modelleme geliştirilebilir.

4.4. Kişilerin Karakter Özellikleri

Bu tez çalışmasında gerçek veri uygulaması bulunmadığından kişilerin karakteristik özellikleri ve sorunun fiziksel yapısı ile ilgili değerlendirme yapılamamış, sadece soru zorluğu ile yanıt süresi arasındaki ilişki kullanılmıştır. Bu durum çalışmanın kısıtlayıcı taraflarından biridir. Literatürde, kişiye özgü özelliklerin (ör. cinsiyet, yaş, ırk, vb.) madde yanıt süresindeki artışı etkileme durumu birçok araştımacı tarafından ele alınmıştır. Çözümsel davranış sergileyen kişilerin madde yanıt süreleri hızlı tahmin davranışı sergileyenlere göre daha uzun olma eğilimindedir. Schmitt ve ark. (1991) yanıt süresi artışında cinsiyetin önemli bir faktör olmadığını ortaya koymuşlardır. Bu çalışmayla tutarlı olarak Cole (1997) yaptığı meta-analitik çalışmada cinsiyetin yanıt süresi gecikmesini etkilemediği sonucuna ulaşmıştır. Yan ve ark. (2015) 50-70 yaşları arasındaki katılımcılarda maddeyi yanıtlamak için harcanan sürenin artışı ile yanıtın kalitesi arasında negatif ilişki bulmuşlardır. Bu sonuca zıt olarak 70 yaş ve üstü katılımcılarda maddeyi yanıtlama süresi ile yanıt kalitesi pozitif ilişkilidir. Yanıt süresi ile etnik köken arasındaki ilişki de literatürde anlamlı bulunmuştur (Schmitt ve Dorans, 1990). Yanıt süresi artışında etkili bulunan bir diğer kişisel özellik, sınava girenlerin kaygı düzeyidir. Bergstrom ve ark. (1994) kaygı düzeyi yüksek olan kişilerin maddeleri doğru yanıtlamak için daha fazla ek zamana ihtiyaç duyduklarını bulmuşlardır. Yanıt süresini etkiliyen ve literatürde dikkat çeken önemli konulardan biri de test sırasında hızlanma durumudur (Mayerl, 2013). Yang ve ark. (2002) çözümsel davranış sergileyen kişilerin, zor maddelere verdikleri yanıt süresinin, hızlı yanıt veren kişilere göre daha fazla olduğunu bulmuşlardır. Ayrıca madde zorluğu ile yanıt

arasında anlamlı bir pozitif ilişki tespit etmişlerdir. Bir başka çalışmada, kişilerin bilgisayar kullanım becerisi arttıkça BUT'larda daha yüksek performans sergiledikleri bildirilmiştir (John, 2013).

4.5. Kişi Sayısı

Literatürde Rasch modelinden farklı olarak, 3PL model ile parametre kestirilirken yüksek sayıda kişiye (yaklaşık olarak n:1000) ihtiyaç duyulduğu belirtilmiştir (De Ayala, 2013). Bu tezde yürütülen benzetim çalışmasında madde parametreleri türetilmiş, ardından gerçek hayatta elimizde yalnızca yanıt deseni olacağı için oluşturulan yanıt deseninden madde parametreleri yeniden kestirilmiştir. Bu işlem gerçekleştirilirken 1000 kişi alınmıştır.

4.6. Çözümsel Davranış ve Hızlı Yanıt Davranışı

Wise ve DeMars (2006) çözümsel davranış modelinin, standart 3PL modelden daha yüksek geçerliğe sahip yetenek düzeyi tahminleri sağlayabileceğini ileri sürmüştür. Bu tez çalışmasında BUT 3PL ile gerçek yetenek düzeyleri arasındaki uyum iyi düzeyde iken, BUT ÇD yöntemiyle elde edilen yetenek düzeyleriyle gerçek yetenek düzeyleri arasındaki uyum orta düzeydeydi. BUT 3PL modelin gerçek yetenek düzeyi ile uyumunun BUT ÇD modele göre daha iyi olduğu görüldü. Gerçek yetenek düzeyinden bağımsız olarak BUT ÇD ve BUT 3PL yöntemleri ile kestirilen yetenek düzeylerinin birbiriyle uyumlu olduğu görüldü.

Literatürde Soland ve ark. (2019) yetenek düzeyi ile hızlı tahmin davranışı (doğal olarak çözümsel davranış) ilişkisine dair yetersiz bilgi olduğunu ifade etmişlerdir. Ayrıca hızlı tahmin oranının da yetenek düzeyi üzerindeki etkisiyle ilgili yeterli kanıt olmayışını dile getirmişlerdi. Shi (2017) çalışmasında BUT uygulamasında yanıt süresi ile yanıt doğruluk oranı arasında negatif ilişki olduğunu tespit etmiş ve bu ilişkinin hem soruyu çok hızlı yanıtlayan (rapid-guessing) hem de soruyu gereğinden çok daha fazla sürede yanıtlayan kişiler için geçerli olduğunu

bildirmiştir. Ayrıca BUT uygulamasında sorulara verilen zaman kısıtının optimum düzeyde olması gerektiğini bildirmiştir. Tipik olarak MYT modelleri, sınava giren kişinin gizli özelliği (latent trait/yetenek düzeyi) arttıkça doğru yanıt olasılığının arttığını varsayar. Dolayısıyla doğru yanıt olasılığı ile kişinin yetenek düzeyi arasında bir ilişki olduğu düşünülmektedir. Bu nedenle Shi'nin çalışmasında belirttiği doğru yanıt olasılığı ile yanıt süresi arasındaki ilişkinin, yetenek düzeyi ile yanıt süresi arasında da geçerli olacağı düşünülebilir. Hızlı tahmin davranışı, sınava giren kişiler soru üzerinde çaba sarfetmeden hızlı bir şekilde yanıtlamaya çalıştıklarında gerçekleşir. Bu tez çalışmasında BUT ÇD yöntemi ile yetenek düzeyi kestirmenin BUT 3PL kestirimlerine göre yetenek düzeyini daha düşük kestirmeye meyilli olduğu görüldü. Hızlı tahmin yapılan durumlarda yani çok kısa sürede soruya yanıt veren kişilerde (ÇD sergilemedikleri maddeler için) yetenek düzeyinin düşük olması beklenen bir durumdu.

5. SONUÇ ve ÖNERİLER

Yanıt süresi kullanımı ile ilgili literatürde birçok model önerilmiştir. Genel olarak madde zorluğu ve kişisel özellikler üzerinde durulan çalışmalarda çok boyutlu etkileşimler irdelenmeden sunulmuştur. Ancak bu sonuçlar kişi yetenek düzeyi kestirimlerinde yanlılık doğurabilir.

Bu tezde yapılan benzetim çalışmasında, yanıt süresi ile madde zorluğu arasındaki korelasyon; çözümsel davranış sergileyen kişiler için pozitif, hızlı yanıt sergileyen kişiler için ise negatif olarak oluşturulmuştur. Bu durumda BUT ÇD ve BUT 3PL ciddi farklara rastlanmamıştır fakat BUT ÇD yöntemi yetenek düzeyini BUT 3PL yöntemine göre biraz daha düşük kestirme eğilimindedir. Klasik BUT uygulamasında olduğu gibi her iki yöntemde de standart hata düzeyinin düşmesi sorulan soru sayısını arttırmıştır ve gerçek yetenek düzeyi ile daha uyumlu kestirimler yapılmasını sağlamıştır. Kâğıt-kalem yöntemleri ile BUT uygulamaları benzer sonuçlar vermiştir. Madde sayısı düşükken (50 madde) BUT uygulamalarında standart hataya (0,3) bağlı algoritma durdurma kriteri sağlanmadığından tüm sorular kişiye sorulmuştur ve ortalama standart hata belirlenen kritere ulaşamamıştır. Fakat madde sayısı artışı ile (100, 250) yöntemlerde durdurma kriteri sağlanmış ve uygulanan soru sayıları azalmıştır.

Kişinin sınav performansını ve davranışını incelemek için daha kısa sürede ve yansız bilgi toplanması adına kâğıt-kalem testleri yerine BUT tercih edilmelidir. Süre kısıtlı BUT'larda kişilerin kaygı düzeyi artacaktır. Buna bağlı olarak kişiler hızlı tahmin davranışına daha çok yönelecektir ve tüm sorulara yanıt alınamayan durumlar oluşabilecektir. Kişilerin çözümsel davranış sergileyebilmesi için, toplam süre kısıtı olan hız testi yerine (speed-test), zaman kısıtı olmayan güç testlerinin (power test) veri toplamak için kullanımı daha uygun olabilir. Eğer güç testi kullanmak mümkün değilse, hız testi kullanılan durumlarda yanıt sürelerini göz önünde bulunduran çözümsel davranış modellerinin kullanılması son derece önemlidir. Fakat uygulanan

testte zaman kısıtı olmaması da bir soruya gereğinden fazla süre harcayıp doğru yanıtlanma oranını düşürebilir.

Bu sonuçlar doğrultusunda çözümsel davranış modeli ile hızlı yanıt modeli arasındaki farklılıkları belirlemek adına; yanıt sürelerinin kullanımı ile ilgili ileri araştırmalar yapılmalı ve bu araştırmalarda, madde zorluğu, kişi yetenek düzeyi, kişinin fiziksel ve bilişsel özellikleri, uygulanan sorunun fiziksel özellikleri eş zamanlı olarak kullanılmalıdır. Benzetim çalışmasıyla ulaşılan sonuçların gerçek ve yüksek kişi sayısına sahip veri setine uygulanması gerekmektedir.

Dijital hastanelerin ve büyük verinin yaygınlaştığı günümüzde sağlık alanında kullanılan ölçekler için de BUT 3PL ve BUT ÇD uygulamaları kullanılabilir. Benzer şekilde içinde bulunduğumuz pandemik krizde uzaktan eğitimin popüler hale gelmesi ile yapılan çevrimiçi sınavlarda kişi yetenek düzeyini daha belirgin şekilde belirlemek ve kişilerin sınav stratejilerini çözümlemek için yanıt süresini içeren BUT ÇD'lerin kullanım alanlarını artacaktır.

BUT ÇD ve BUT 3PL arasında benzetim çalışması sonucunda oluşan farklılıklar gerçek veri setinde farklılaşabilir. Ayrıca modelde soru seçimi için bilgi kriteri olarak verimlilik dengeli (Efficiency Balanced) bilgi kriteri, Ağırlıklandırılmış olabilirlik (Weighted Likelihood) bilgi kriteri, zorluk eşleştirmeli (Difficulty Matching) bilgi kriteri ve Kullback-Leibler bilgi kriteri gibi farklı kriterler kullanılabilir. Zamana bağlı eşik değerin belirlenmesi için kullanılan yöntem ve BUT'a katılan kişi sayısı etkili faktörler olabilir, bu yüzden konuyla ilgili daha fazla çalışmaya gerek vardır. Hızlı tahmin yönteminde süreye ait eşik değeri kullanarak yetenek düzeyini kestiren ÇD modeline ek olarak çok uzun süreli yanıtlar için de farklı bir modelleme geliştirilebilir. Benzer şekilde BUT algoritmasını durdurma kriteri olarak standart hata dışında madde sayısı ile durdurma kriteri de kullanılabilir.

ÖZET

Bilgisayar Uyarlamalı Testlerde Madde Yanıt Sürelerinin Kullanımı

Teknoloji ve bilgi çağındaki gelişmelere paralel olarak kâğıt-kalem testleri yerini Bilgisayar Uyarlamalı Testlere (BUT) bırakmaktadır. Gelişen bilgisayar teknolojisi yardımıyla bu testlerde yalnızca kişinin yanıt deseni ile ilgili bilgi elde edilmez, bu duruma ek olarak yanıt süreleri ve kişinin maddeyi yanıtladığı andaki davranış şekli ile ilgili bilgi toplanabilir. Yanıt süresi kullanımı ile ilgili literatürde birçok model önerilmiştir. Genel olarak madde zorluğu ve kişisel özellikler üzerinde durulan çalışmalarda çok boyutlu etkileşimler irdelenmeden sunulmuştur ve bu sonuçlar kişi yetenek düzeyi kestirimlerinde yanlılık doğurabilir.

Bu çalışmanın amacı, madde yanıt teorisi temelli üç parametreli lojistik (3PL) model ile oluşturulan BUT ile yanıt süresini içeren çözümsel davranış (ÇD) modeli temelli BUT uygulamalarının performanslarının karşılaştırılmasıdır. Ayrıca BUT ÇD ve BUT 3PL modeller dışında kâğıt-kalem ÇD ve kâğıt-kalem 3PL modellerin birbirleriyle ve BUT ile uyumlarını değerlendirmektir. Bu amaç doğrultusunda planlanan benzetim çalışmasında oluşturulan madde parametrelerinden zorluk parametresi 0 ortalama ve 1,5 standart sapmalı normal dağılımdan N(0,1.5); ayırt edicilik parametresi tekdüze dağılımdan U(0,2) türetilmiştir. Üç parametreli lojistik modelde yer alan ve alt asimptot olarak da isimlendirilen şansa bağlı tahmin parametresi ise 5 yanıtlı sınav sistemi göz önünde bulundurularak U(0.2,0.5) dağılımdan türetilmiştir. Kullanılan ÇD ve 3PL modellere BUT uygulaması yapılırken BUT sonlanma kriteri standart hatanın <0,3 ve <0,5 olması şeklinde seçilmiştir. Kişi sayısı 1000 olarak belirlenirken, soru bankasındaki madde sayısı 50, 100 ve 250 olarak değişmektedir. Benzetim çalışmasında kullanılan 6 senaryoda tüm işlemler 1000 kez tekrar edilmiştir.

Sonuç olarak, soru bankasındaki madde sayısının artışı ve standart hatanın azalması BUT ÇD ve BUT 3PL için birçok durumda gerçek yetenek düzeyi ile uyumlu kestirimler yapılmasını sağlamıştır. BUT ÇD yöntemi yetenek düzeyini BUT 3PL'ye göre önemsiz derecede daha düşük kestirmektedir. BUT uygulamasında ve kâğıt-kalem yöntemlerinde yapılan kestirimler hem çözümsel davranışta hem de klasik 3PL modelde uyumlu bulunmuştur.

Anahtar sözcükler: Bilgisayar uyarlamalı test, madde yanıt teorisi, üç parametreli lojistik model, çözümsel davranış modeli, yanıt süresi

SUMMARY

Use of the Item Response Times in Computerized Adaptive Testing

Depending on the developments in technology and information, paper-pencil tests leave their place to Computerized Adaptive Tests (CAT). With the help of developing computer technology, information about the response pattern of the person is not only obtained in these tests, in addition to this situation, information about the response times and the behavior of the person in the test could be collected. Many models have been proposed in the literature regarding the use of response time. These studies were focused on item difficulty and personal characteristics by ignoring the multidimensional interactions, therefore these results may cause bias in estimates of individual ability levels.

The aim of this study is to compare the performance of CAT applications of classical three-parameter logistic (3PL) model which based on the item response theory and solution behavior (SB) model that based on response time. In addition, to evaluate the agreement of paper-pencil SB and paper-pencil 3PL models with each other and with CAT applications. In the simulation study planned for this purpose, difficulty parameter (b) was derived from normal distribution with zero mean and one and half standard deviations N(0, 1.5) and discrimination parameter (a) was derived from uniform distribution U(0,2). Guessing parameter (c), which is included in the three-parameter logistic model and also named as lower-asymptote was derived from uniform distribution U(0.2,0.5) by taking the 5-answer exam system into consideration. While applying CAT to the SB and 3PL models, the standard error in the CAT was taken as <0.5 and <0.3 as stopping criteria. The number of person were determined as 1000, the number of items in the question bank were changed as 50, 100 and 250. All 6 scenarios which were used in this simulation study were repeated 1000 times.

With the increase in the number of items in the question bank and the decrease in the standard error, consistent results were obtained with true ability level in the CAT SB and CAT 3PL methods. The CAT SB method estimates the ability level slightly lower than CAT 3PL. The estimations which were made in CAT application and paper and pencil methods were found to be consistent in both SB and classical 3PL models.

Keywords: Computerized adaptive test, item response theory, three parameter logistic model, solution behavior model, response time

KAYNAKLAR

BENNETT RE, BRASWELl J, ORANJE A, SANDENE B, KAPLAN B, YAN F (2008). Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, **6(9)**:1–39.

BERGSTROM BA, GERSHON RC, LUNZ, ME (1994). "Computer adaptive testing: exploring examinee response time using hierarchical linear modeling," Paper Presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA).

BRIDGEMAN B, CLINE F (2000). Varıatıons In Mean Response Tımes For Questions On The Computer- Adaptıve GRE® General Test: Implicatıons For Fair Assessment. *ETS Research Report Series*, **2000(1)**:i-29.

CALLEAR D, KING T (1997). Using computer-based tests for information science. ALT-J, **5(1)**:27-32.

CHAE YM, PARK SG, PARK I (2019). The relationship between classical item characteristics and item response time on computer-based testing. *Korean journal of medical education*, **31(1)**:1.

CHANG SW, TWU BY (1998). A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing. Erişim Adresi: [https://files.eric.ed.gov/fulltext/ED420722.pdf]. Erişim tarihi: 09/11/2020.

CHOI SW (2009). Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT Models (Computer software). *Applied Psychological Measurement*, **33**:644-645.

COLE NS (1997). The ETS Gender Study: How Females and Males Perform in Educational Settings. New Jersey, NJ: Educational Testing Service)

DE AYALA RJ (2013). The theory and practice of item response theory. Guilford Publications.)

ELHAN AH (2002). Rasch analizinin incelenmesi ve fiziksel tıp ve rehabilitasyon alanında bir uygulaması (Doctoral dissertation, Doktora Tezi, Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü, Ankara).

ELHAN AH, ATAKURT Y (2005). Ölçeklerin değerlendirilmesinde niçin Rasch analizi kullanılmalıdır?. Ankara Üniversitesi Tıp Fakültesi Mecmuası, **58(01)**:47-50.

GAMER M, LEMON J, GAMER MM, ROBINSON A, KENDALL'S W (2012). Package 'irr'. Various coefficients of interrater reliability and agreement. CRAN R Project.

HALKITIS PN, JONES P, PRADHAN J (1996). Estimating testing time: the effects of itemcharacteristics on response latency. *Paper Presented at the Annual Meeting of the American Educational Research Association* (New York, NY).

HOL AM, VORST HC, MELLENBERGH GJ (2008). Computerized adaptive testing of personality traits. *Zeitschrift für Psychologie/Journal of Psychology*, **216(1)**:12-21.

JOHN SP (2013). Influence of computer self-efficacy on information technology adoption. *International Journal of Information Technology*, **19(1)**:1-13.

LAWRENCE IM (1993). The Effect of Test Speededness on Subgroup Performance. New Jersey, NJ: Educational Testing Service. ERIC document #386494)

LEE YH, CHEN H (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, **53(3)**:359.

LEHNERT B (2015). BlandAltmanLeh: plots (slightly extended) Bland-Altman plots. R package version 0.3,1.

MADSEN HS (1991). Computer-adaptive testing of listening and reading comprehension. (In P. Dunkel Ed.), Computer-assisted language learning and testing: Research issues and practice (p:237-257). New York: Newbury House.

MASTERS J, SCHNIPKE DL, CONNOR C (2005). "Comparing item response times and difficulty for calculation items," Paper Presented at the Annual Meeting of the American Educational Research Association (Montréal, QC: Canada)

MAYERL J (2013). "Response latency measurement in surveys. Detecting strong attitudes and response effects," in Survey Methods: Insights from the Field. Erişim Adresi: [http://surveyinsights.org/?p=1063]. Erişim Tarihi: 05/11/2019.

PARSHALL CG (1994). Response Latency: An Investigation into Determinants of Item-Level Timing. Erişim Adresi: [https://files.eric.ed.gov/fulltext/ED370961.pdf]. Erişim Tarihi: 01/07/2020.

PARTCHEV I (2016). Package irtoys: Simple interface to the estimation and plotting of IRT models. CRAN R Project.

REEVE BB (2002). An introduction to modern measurement theory. National Cancer Institute, 1-67. Erişim Adresi: [http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.207.3244&rep=rep1&type=pdf], Erişim Tarihi: 14/12/2019.

REZAIE M, GOLSHAN M (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, **2(5)**:128-137.

RUDNER LM (1998). An online, interactive, computer adaptive testing tutorial. 11/98. Erişim Adresi: [http://edres.org/scripts/cat/catdemo.htm]. Erişim tarihi: 05/11/2019.

SALKIND NJ (Ed.) (2010). Encyclopedia of research design **(Vol. 1)**. Thousand Oaks, CA: SAGE Publications, Inc.

SCHMITT AP, BLEISTEIN CA (1987). Factors affecting differential item functioning for black examinees on scholastic aptitude test analogy items 1. *ETS Research Report Series*, **1987(1)**:i-46.

SCHMITT AP, DORANS NJ (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, **27(1)**:67-81.

SCHMITT AP, DORANS NJ, CRONE CR, MANECKSHANA BT (1991). Differential Speededness and Item Omit Patterns on the SAT (Research Report No 91-50). Princeton, NJ: Educational Testing Service.

SCHNIPKE DL (1999). The influence of speededness on item-parameter estimation (Computerized Testing Report No. 96-07). Princeton, NJ: Law School Admission Council (ERIC Document Reproduction Service No. ED467809). Erişim Adresi:[https://files.eric.ed.gov/fulltext/ED467809.pdf]. Erişim tarihi: 07/12/2019.

SCHNIPKE DL, SCRAMS DJ (1997). Modeling item response times with a two- state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, **34(3)**:213-232.

SCRAMS DJ, SCHNIPKE DL (1997). Making Use of Response Times in Standardized Tests: Are Accuracy and Speed Measuring the Same Thing?. Erişim Adresi: [https://files.eric.ed.gov/fulltext/ED409357.pdf]. Erişim Tarihi: 02/07/2020.

SHI Y (2017). Response Time and Response Accuracy in Computerized Adaptive Testing. Erişim Adresi:[https://iacat.org/response-time-and-response-accuracy-computerized-adaptive-testing]. Erişim tarihi:15/09/2020.

SI F, SCHUMACKER RE (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing*, **4(2)**:137-181.

SMITH RW (2000). "An exploratory analysis of item parameters and characteristics that influence itemresponse time," Paper Presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA.)

SOLAND J, WISE SL, GAO L (2019). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education*, **32(2)**: 151-165.

THOMPSON NA, WEISS DJ (2011). A framework for the development of computerized adaptive tests. *Pract Assess Res Eval*. **16**:1–9.

THRELFALL J, POOL P, HOMER M, SWINNERTON B (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. *Educational Studies in Mathematics*, **66(3)**:335–348.

TURGUT MF (1997). Eğitimde ölçme ve değerlendirme metotları (10. Baskı). Ankara: Gül Yayınevi.

VAN DER LINDEN WJ (2006). A lognormal model for response times on test items. *J. Educ. Behav.* Stat. **31**:181–204.

VAN DER LINDEN WJ, VAN KRIMPEN-STOOP EMLA (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika* **68**:251–265

WAINER H (1990). Computerized Adaptive Testing: A primer. Hillsdale, N J: LEA.

WAINER H, DORANS NJ, FLAUGHER R, GREEN BF, MISLEVY RJ (2000). Computerized adaptive testing: A primer. Routledge.

WANG S, JIAO H, YOUNG M, BROOKS T, OLSON J (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement*, **68**:5–24.

WANG T, HANSON BA (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, **29(5)**: 323-339.

WEISS DJ, KINGSBURY GG (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, **21(4)**:361-375.

WISE, SL (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, **19**:25-114.

WISE SL, DEMARS CE (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, **43(1)**:19-38.

WISE SL, KONG X (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, **18(2)**:163-183.

WISE SL, MA L (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada.

YAN T, RYAN L, BECKER SE, SMITH J. (2015). Assessing quality of answers to a global subjective well-being question through response times. *Sur. Res. Meth*ods, **9**:101–109.

YANG CL, O'NEILL TR, KRAMER GA (2002). Examining item difficulty and response time on perceptual ability test items. *J. Appl. Measure*. **3**:282–299.

YOUNG R, SHERMIS MD, BRUTTEN SR, PERKINS K (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System*, **24(1)**:23-40.

EKLER

Ek-1. Benzetim çalışmasında kullanılan kodların bazıları

(soru sayısı:50, SE_{CAT}:0,5)

##Kullanılan Paketlerin Yüklenmesi

```
install.packages("irr")
install.packages("ggExtra")
library("irr")
library(ggExtra)
```

Benzetim Tekrarı

```
nrepeat=1000
```

#Kişi sayısı

```
nSimulee<-1000
```

#Madde sayısı

```
ni=50
```

#Diğer parametreler

```
rho.p=0.6, rho.n=-0.6, popMean=0, popSD=1,5, maxCat=2, minTheta=-4, maxTheta=4, inc=0.1, maxNI=ni, minNI=1, maxSE=0.5, topN=1, first.item.selection=1, first.item=1, prior.mean=0, prior.sd=1, time.mean1=6, time.mean2=60, time.sd1=4, time.sd2=20

theta=seq(minTheta,maxTheta,inc), nq=length(theta), start.theta=prior.mean, prior=dnorm((theta-prior.mean)/prior.sd)
```

#Sonuçlara ait matris oluşturma

```
BUT.Results=data.frame(matrix(NA,nrepeat,8))
Items.Par=data.frame(matrix(NA,ni,nrepeat*2))
Items.Freq=data.frame(matrix(NA,ni,nrepeat*2))
All.Theta=data.frame(matrix(NA,nrepeat*nSimulee,5))
ICC=data.frame(matrix(NA,nrepeat,4))
Response.time=data.frame(matrix(NA,nrepeat*nSimulee,ni))
```

##Benzetim Tekrarı

```
cut=1

repeat{
```

##Madde parametrelerinin türetilmesi

```
a=runif(ni,0,2)
cb=rnorm(ni,0,1.5)
g=runif(ni,0.2,0.5)
NCAT=rep(2,ni)
item.par=data.frame(t(rbind(a,cb,g,NCAT)))
```

#Madde zorluğu ile korele sürenin türetilmesi

```
genTime=function(n,ni){
  time=matrix(0,n,ni)
  if(ni==50){
    for (i in 1:n*0.1){
      theta=acos(rho.n)
      x2=rnorm(ni,6,4)
      X=cbind(cb, x2)
      Xctr=scale(X, center=TRUE, scale=FALSE)
      Id=diag(ni)
      Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
      P=tcrossprod(Q)
      x2o=(Id-P) %*% Xctr[ , 2]
      Xc2=cbind(Xctr[ , 1], x2o)
      Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
      x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
      x=x*22.4+6
      time[i,]=x
    }
    for (i in (n*0.1+1):n){
      theta=acos(rho.p)
      x2=rnorm(ni,60,20)
      X=cbind(cb, x2)
      Xctr=scale(X, center=TRUE, scale=FALSE)
      Id=diag(ni)
      Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
      P=tcrossprod(Q)
      x2o=(Id-P) %*% Xctr[ , 2]
      Xc2=cbind(Xctr[ , 1], x2o)
      Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
      x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
      x=x*112+60
      time[i,]=x
    }
  } else if(ni==100){
    for (i in 1:n*0.1){
      theta=acos(rho.n)
      x2=rnorm(ni,6,4)
      X=cbind(cb, x2)
      Xctr=scale(X, center=TRUE, scale=FALSE)
      Id=diag(ni)
      Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
      P=tcrossprod(Q)
      x2o=(Id-P) %*% Xctr[ , 2]
      Xc2=cbind(Xctr[ , 1], x2o)
```

```
    Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
    x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
    x=x*31.8396+6
    time[i,]=x
   }
   for (i in (n*0.1+1):n){
    theta=acos(rho.p)
    x2=rnorm(ni,60,20)
    X=cbind(cb, x2)
    Xctr=scale(X, center=TRUE, scale=FALSE)
    Id=diag(ni)
    Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
    P=tcrossprod(Q)
    x2o=(Id-P) %*% Xctr[ , 2]
    Xc2=cbind(Xctr[ , 1], x2o)
    Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
    x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
    x=x*159.19798+60
    time[i,]=x
   }
 } else if(ni==250){
   for (i in 1:n*0.1){
    theta=acos(rho.n)
    x2=rnorm(ni,6,4)
    X=cbind(cb, x2)
    Xctr=scale(X, center=TRUE, scale=FALSE)
    Id=diag(ni)
    Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
    P=tcrossprod(Q)
    x2o=(Id-P) %*% Xctr[ , 2]
    Xc2=cbind(Xctr[ , 1], x2o)
    Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
    x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
    x=x*50.49515+6
    time[i,]=x
   }
   for (i in n*0.1+1:n){
    theta=acos(rho.p)
    x2=rnorm(ni, 60, 20)
    X=cbind(cb, x2)
    Xctr=scale(X, center=TRUE, scale=FALSE)
    Id=diag(ni)
    Q=qr.Q(qr(Xctr[ , 1, drop=FALSE]))
    P=tcrossprod(Q)
    x2o=(Id-P) %*% Xctr[ , 2]
    Xc2=cbind(Xctr[ , 1], x2o)
    Y=Xc2 %*% diag(1/sqrt(colSums(Xc2^2)))
    x= Y[ , 2] + (1 / tan(theta)) * Y[ , 1]
    x=x*252.47575+60
    time[i,]=x
   }
 }
time=data.frame(time)
```

```
  return(time)
 }

 resp.time=genTime(nSimulee,ni)
 names(resp.time)<-c(paste("T",1:ni,sep=""))
```

#Yanıt deseninin türetilmesi

Bu kodlar için yazar ile iletişime geçiniz

#Kullanılan maddeler ve BUT sonuçlar için matrislerin oluşturulması

```
items.used<-matrix(NA,nExaminees,maxNI)
 selected.item.resp<-matrix(NA,nExaminees,maxNI)
 ni.administered<-numeric(nExaminees)
 theta.CAT<-rep(NA,nExaminees)
 sem.CAT<-rep(NA,nExaminees)
 theta.history.CAT<-matrix(NA,nExaminees,maxNI)
 se.history.CAT<-matrix(NA,nExaminees,maxNI)

 theta.3PL<-rep(NA,nExaminees)
 sem.3PL<-rep(NA,nExaminees)

 items.used.SB<-matrix(NA,nExaminees,maxNI)
 selected.item.resp.SB<-matrix(NA,nExaminees,maxNI)
 ni.administered.SB<-numeric(nExaminees)
 theta.SB<-rep(NA,nExaminees)
 sem.SB<-rep(NA,nExaminees)
 theta.history.SB<-matrix(NA,nExaminees,maxNI)
 se.history.SB<-matrix(NA,nExaminees,maxNI)

 theta.3PL.SB<-rep(NA,nExaminees)
 sem.3PL.SB<-rep(NA,nExaminees)

 NCAT=item.par[,"NCAT"]
 DISC=item.par[,"a"]
 CB=item.par[,"cb"]
 GUESS=item.par[,"g"]
```

#information hesabı için hazırlık matrisinin oluşturulması

Bu kodlar için yazar ile iletişime geçiniz

#information hesaplanması (BUT 3PL ve BUT ÇD)

Bu kodlar için yazar ile iletişime geçiniz

#Süreyle ilgili eşik değerin ÇD için kullanılması

```
        threshold=ifelse(mean(resp.time[,item])*0.1>10,10,mean(resp.time[,item])*0.1)

        if (i==1){
          pstar[2]=ifelse(resp.time[examinee,item]>threshold,g+(1-g)/(1+exp(-a*(pre.theta-
cb))),g+0.001)
        } else {
          pstar[2]=ifelse(resp.time[examinee,item]>threshold,g+(1-g)/(1+exp(-a*(pre.theta-
cb))),g)
        }

        for (k in 1:ncat[i]) {
          pp1[k]=pstar[k]-pstar[k+1]
          qq1[k]=1-pp1[k]
        }
        if(resp[i]==1){
          deriv1=deriv1-a/(1-g)*(pp1[2]-g)
          deriv2=deriv2-(a/(1-g))^2*(pp1[2]-g)*(qq1[2]-pp1[2]+g)-(a/(1-g)*(pp1[2]-g))^2
        }  else {
          deriv1=deriv1+a/(1-g)*(pp1[2]-g)*qq1[2]/pp1[2]
          deriv2=deriv2+(a/(1-g))^2*((pp1[2]-g)*qq1[2]*(qq1[2]-pp1[2]+g)/pp1[2]-((pp1[2]-
g)*qq1[2]/pp1[2])^2)
        }
       }
```

 # BUT uygulaması

```
 for (j in 1:nExaminees) {

   ##standard CAT
   critMet<-FALSE
   items.available<-rep(TRUE,ni)
   items.available[is.na(resp.data[j,paste("R",1:ni,sep="")])]<-FALSE
   max.to.administer<-ifelse(sum(items.available)<=maxNI,sum(items.available),maxNI)
   ni.given<-0
   theta.current<-start.theta

   while (critMet==FALSE && ni.given<max.to.administer) {
     array.info<-calcInfo(theta.current)
     ni.available<-sum(array.info>0)
     info.index<-rev(order(array.info))
     item.selected<-select.maxInfo()
     ni.given<-ni.given+1
     items.used[j,ni.given]<-item.selected
     items.available[item.selected]<-FALSE
     selected.item.resp[j,ni.given]<-resp.data[j,paste("R",item.selected,sep="")]
     estimates<-calcMLE(j,ni.given)

     theta.history.CAT[j,ni.given]<-estimates$THETA
     se.history.CAT[j,ni.given]<-estimates$SEM
     theta.current<-estimates$THETA
```

```
    if (ni.given>=max.to.administer || (estimates$SEM<=maxSE && ni.given>=minNI)) {
      critMet<-TRUE
      theta.CAT[j]<-estimates$THETA
      sem.CAT[j]<-estimates$SEM
      ni.administered[j]<-ni.given
    }
  }

  ##standard 3PL model
  estimates.3PL<-calcMLE.3PL(j,ni)
  theta.3PL[j]<-estimates.3PL$THETA
  sem.3PL[j]<-estimates.3PL$SEM


  ##SB CAT
  critMet.SB<-FALSE
  items.available.SB<-rep(TRUE,ni)
  items.available.SB[is.na(resp.data[j,paste("R",1:ni,sep="")])]<-FALSE
  max.to.administer.SB<-
ifelse(sum(items.available.SB)<=maxNI,sum(items.available.SB),maxNI)
  ni.given.SB<-0
  theta.current.SB<-start.theta

  while (critMet.SB==FALSE && ni.given.SB<max.to.administer.SB) {
    array.info.SB<-calcInfo.SB(theta.current.SB)
    ni.available.SB<-sum(array.info.SB>0)
    info.index.SB<-rev(order(array.info.SB))
    item.selected.SB<-select.maxInfo.SB()
    ni.given.SB<-ni.given.SB+1
    items.used.SB[j,ni.given.SB]<-item.selected.SB
    items.available.SB[item.selected.SB]<-FALSE
    selected.item.resp.SB[j,ni.given.SB]<-resp.data[j,paste("R",item.selected.SB,sep="")]
    estimates.SB<-calcMLE.SB(j,ni.given.SB)

    theta.history.SB[j,ni.given.SB]<-estimates.SB$THETA
    se.history.SB[j,ni.given.SB]<-estimates.SB$SEM
    theta.current.SB<-estimates.SB$THETA

    if (ni.given.SB>=max.to.administer.SB|| (estimates.SB$SEM<=maxSE &&
ni.given.SB>=minNI)) {
      critMet.SB<-TRUE
      theta.SB[j]<-estimates.SB$THETA
      sem.SB[j]<-estimates.SB$SEM
      ni.administered.SB[j]<-ni.given.SB
    }
  }

  ##SB 3PL model
  estimates.3PL.SB<-calcMLE.3PL.SB(j,ni)
  theta.3PL.SB[j]<-estimates.3PL.SB$THETA
  sem.3PL.SB[j]<-estimates.3PL.SB$SEM
}
```

##madde parametrelerinin yazdırılması

```
Items.Par[ ,c(2*cut-1,2*cut)]=item.par[,c(1,2)]
```

##SKK'nın yazdırılması

```
ICC[cut,1]=(icc(cbind(resp.data$theta,theta.CAT),model="twoway",type="agreement",unit="single"))$value

ICC[cut,2]=(icc(cbind(resp.data$theta,theta.SB),model="twoway",type="agreement",unit="single"))$value

ICC[cut,3]=(icc(cbind(resp.data$theta,theta.3PL),model="twoway",type="agreement",unit="single"))$value

ICC[cut,4]=(icc(cbind(resp.data$theta,theta.3PL.SB),model="twoway",type="agreement",unit="single"))$value
```

#thetaların yazdırılması

```
All.Theta[c(1:nSimulee)+(cut-1)*nSimulee,1]=resp.data$theta
All.Theta[c(1:nSimulee)+(cut-1)*nSimulee,2]=theta.CAT
All.Theta[c(1:nSimulee)+(cut-1)*nSimulee,3]=theta.SB
All.Theta[c(1:nSimulee)+(cut-1)*nSimulee,4]=theta.3PL
All.Theta[c(1:nSimulee)+(cut-1)*nSimulee,5]=theta.3PL.SB
```

#yanıt süresinin yazdırılması

```
Response.time[c(1:nSimulee)+(cut-1)*nSimulee,c(1:ni)]=resp.time[,c(1:ni)]
```

#ortalama madde sayısının yazdırılması

```
BUT.Results[cut,1]=sum(!is.na(items.used))/nrow(resp.data)
#median number of items used in CAT
n.items.used.CAT=matrix(0,nSimulee,1)

for(i in 1:nExaminees){
  n.items.used.CAT[i,1]=sum(!is.na(items.used[i,]))
}
BUT.Results[cut,2]=median(n.items.used.CAT)
#mean SE of CAT
BUT.Results[cut,3]=mean(sem.CAT)
#mean bias of CAT
BUT.Results[cut,4]=mean(resp.data$theta-theta.CAT)
#average number of items used in SB CAT
BUT.Results[cut,5]=sum(!is.na(items.used.SB))/nrow(resp.data)
#median number of items used in CAT

n.items.used.SB=matrix(0,nSimulee,1)

for(i in 1:nExaminees){
  n.items.used.SB[i,1]=sum(!is.na(items.used.SB[i,]))
```

```
 }
 BUT.Results[cut,6]=median(n.items.used.SB)
 #mean SE of SB CAT
 BUT.Results[cut,7]=mean(sem.SB)
 #mean bias of CAT
 BUT.Results[cut,8]=mean(resp.data$theta-theta.SB)

 cut=cut+1
 if(cut>nrepeat) break
}

names(BUT.Results)=c("Mean.item","Median.item","Sem.CAT","Mean.Bias.Theta","Mean.
item.SB","Median.item.SB","Sem.SB","Mean.Bias.Theta.SB")
names(All.Theta)=c("True","CAT","SB.CAT","ThreePL","SB")
names(ICC)=c("ICC.CAT","ICC.SB.CAT","ICC.3PL","ICC.SB.3PL")

##names(Items.Par)=c(paste(c("a","b"),1:nrepeat,sep=""))
```

#Dosyaların dışa aktarımı

```
write.table(All.Theta, file = "All.Theta.txt", row.names = FALSE, sep = "  ")
write.table(Items.Par, file = "Items.Par.txt", row.names = FALSE, col.names = FALSE, sep
= "  ")
write.table(BUT.Results, file = "BUT.Results.txt", row.names = FALSE, col.names =
FALSE, sep = "  " )
write.table(ICC, file = "ICC.txt", row.names = FALSE, col.names = FALSE, sep = "  " )
write.table(Response.time, file = "Response.time.txt", row.names = FALSE, col.names =
FALSE, sep = "  " )
```