Investigation of Measurement Precision and Test Length in Computerized Adaptive Testing Under Different Conditions

Res. Assist. Dr. Ebru Balta
Ağrı Ibrahim Cecen University - Turkey
ORCID: 0000-0002-2173-7189
ebrubalta2@gmail.com

Assist. Prof. Dr. Arzu Uçar
Hakkari University - Turkey
ORCID: 0000-0002-0099-1348
arzukapcik@gmail.com

Abstract

Computerized Adaptive Tests (CAT) are gaining much more attention than ever by the institutions especially the ones attracting students worldwide due to the nature of CAT not allowing the same items to be presented to different individuals taking the test. In this study, it was aimed to investigate of measurement precision and test length in computerized adaptive testing (CAT) under different conditions. The research was implemented as a Monte Carlo simulation study. In line with the purpose of the study, 500 items which response probabilities were modeled with the three parameter logistic (3PL) model were generated. Fixed length (15,20), standard error (SE<.30, SE<.50) termination rules have been used for the study. Additionally, in comparing termination rules, different starting rules (θ=0,-1<θ< 1), ability estimation methods (Maksimum Likelihood Estimation (MLE) ,Expected a Posteriori (EAP) and Maximum a Posteriori Probability (MAP)), item selection method (Kullback Leibler Information (KLI) and Maximum Fischer Information (MFI)) have been selected since these are critical in the algorithms of CAT. 25 replications was performed for each condition in the generated data. The results obtained from study were evaluated by using RMSE, bias and fidelity values criterions. R software was used for data generation and analyses. As a result of the study, it was seen that choosing the test starting rule as θ=0 or -1<θ<1 did not cause a significant difference in terms of measurement precision and test length. It was concluded that the termination rule, in which RMSE and bias values were lower than the other conditions, was the 0.30 SE termination rule. When the EAP ability estimation method was used, lower RMSE and bias values were obtained compared to the MLE. It was concluded that the KLI item selection method had lower RMSE and bias values compared to the MFI.

Keywords: *Computerized adaptive testing, Item response theory, Measurement precision, RMSE*

E-International Journal of Educational Research

Vol: 13, No: 1, pp. 51-68

Research Article

Received: 2021-11-13
Accepted: 2022-01-26

Suggested Citation

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

Extended Abstract

Problem: The application of large-scale and high-risk exams, which have an important place in making important decisions in the lives of individuals, and the necessity of online exam applications used in the distance education process, measurement and evaluation applications are gaining momentum today with the global COVID 19 pandemic period. Computerized Adaptive Testing (CAT) are gaining much more attention than ever by the institutions especially the ones attracting students worldwide due to the nature of CAT not allowing the same items to be presented to different individuals taking the test. CAT applications consists of the creation of the item pool, the process of starting the test, ability estimation methods, item selection methods and the test termination rule (Kingsbury & Zara, 1989; Orcutt, 2002; Wainer, 2000).

In CAT algorithm, firstly, the ability level of the individual is tried to be estimated. In determining the difficulty level of the first item chosen at the beginning of the test, prior knowledge about the individual's ability can be used, and it is stated that in the absence of prior knowledge about the individual's ability, medium difficulty level item use will be more effective psychometrically In determining the difficulty level of the first item chosen at the beginning of the test, prior knowledge about the individual's ability can be used, and it is stated that in the absence of prior knowledge about the individual's ability, medium difficulty level item use will be more effective psychometrically (Mills & Stocking, 1996; Sereci, 2003). After the selection of the first item for estimation individual's ability, Maximum Likelihood Estimation method (MLE; Birnbaum, 1968), Weighted Likelihood Estimation method (WLE; Warm, 1989), Marginal Maximum Likelihood Estimation method (MMLE) and based on bayesian methods as Expected Posterior Distribution method (EAP), Maximum Posterior Distribution method (MAP) are frequently used (Baker & Kim, 2004; Embretson & Reise, 2000). In CAT applications, the next step after estimating the initial ability levels of the individuals is the selection of the items to be applied to the individual in the continuation of the test. The simplest item selection algorithm for CAT is random item selection. Random item selection is not considered an effective method because it is performed without using information about the items and individuals. As a more effective approach, an intelligent item selection approach has been developed, based on cut-off points and prediction-based. The items that give the highest information about the cut-off point are selected atmaximum discrimination, information gain and minimum expected value based on Classical Test Theory (CTT) , Maximum Fisher Information (MFI), Kullback Leibler Information (KLI) based on Item Response Theory (IRT) and log-odds ratio cut-off point-based methods (Thompson, 2007b). There are two different test termination rules, fixed and variable test length, which have different effects on the bias of ability estimations and the standard error of the test, differing according to the purpose of the test and the characteristics of the item pool (Babcock & Weiss, 2012; Blais & Raiche, 2010; Segall, 2004; Sereci, 2003; Weiss & Kingsbury, 1984). In this study, it was aimed to investigate of measurement precision and test length in CAT under different conditions.

Method: Monte Carlo simulation study was conducted to investigate of measurement precision and test length in CAT under different conditions. The data generation of the computerized adaptive test was carried out using *"catR"* package in R programme. In line with the purpose of the study, 500 items which the response probabilities were modeled with the three parameter logistic (3PL) model were generated.The distribution of (a) discrimination ,(b) difficulty, and (c) guessing parameters of the items in the CAT pool have the following minimum and maximum: (a) min .40, max 1.61; (b) min-3, max 3.8; (c) min .0, max .24. There were 1000 nonaberrant examinees were simulated with abilities drawn from N(0, 1). In comparing termination rules, different starting rules (θ=0, -1<θ<1), ability estimation methods (Maksimum Likelihood Estimation (MLE), Maximum a Posteriori Probability (MAP) and Expected a Posteriori (EAP)), item selection method (Kullback-Leibler Information (KLI) and Maximum Fischer Information (MFI)), termination rules (fixed length (15, 20), standard error (SE<.30, SE<.50) have been selected since these are critical in the algorithms of CAT. 25 replications was performed for each condition in the generated data. The results obtained from study were evaluated by using RMSE, bias and fidelity values criterion. Analyses was carried out using the codes written by the researchers and the "*catR*" package in of R programme.

Findings: Among all the conditions discussed, the lowest RMSE value was obtained as 0.296 when the starting rule was zero (θ=0), the item selection method was KLI, the ability estimation method was MLE, and the test termination rule was SE<0.30. In addition, the same lowest RMSE value was observed when the starting rule was -1<θ<1, the item selection method was KLI, the ability estimation method was EAP, and the test termination rule was SE<0.30. Under all conditions, the highest RMSE value was obtained with 0.86 when the starting rule was zero (θ=0), item selection method was MFI, ability estimation method was MLE, and the test termination rule was fixed length with 15 items. It was observed that the lowest bias value was 0 under all conditions and the rule for starting the test was zero (θ=0), the item selection method was KLI, the ability estimation method was MAP, and the test termination rule was SE<0.30. The highest bias value was obtained as 0.056 under all conditions. This value was obtained under the condition that the starting rule was zero (θ=0), the item selection method was MFI, the ability estimation method was MLE, and the test termination rule was SE<0.50. The highest mean fidelity value was observed as 0.96 under the condition that the starting rule was zero (θ=0), the item selection method was KLI, the ability estimation method was MLE, and the test termination rule was SE<0.30. The lowest mean fidelity value was 0.721 in the condition of zero (θ=0) for the starting rule, MFI, for the item selection method, MAP, for the ability estimation method, and SE<0.50 for the test termination rule. When the average test lengths obtained for each condition are examined, It was observed that the test was terminated with least 8 items, the starting rule is zero (θ=0), the item selection method is MFI, the ability estimation method is MAP and the test termination rule is SE<0.50, and the starting rule is -1<θ< 1, it was observed that the item selection method is MFI, the ability estimation method is MAP, and the test termination rule SE<0.30. It was observed that the test was terminated with the most 40 items when the starting rule was zero (θ=0), the item selection method was MFI, the ability estimation method was MLE, and the test termination rule was SE<0.30.

Suggestions: A similar study can be performed with different item pool sizes. In addition, by changing the properties of the item pool, termination rules can be compared. Item exposure rate was not taken into account in the study. Similar studies that take into account the rate of item exposure can be carried out.

Bilgisayar Ortamında Bireye Uyarlanmış Test Uygulamalarında Ölçme Kesinliğinin ve Test Uzunluğunun Farklı Koşullar Altında İncelenmesi[1]

Arş. Gör. Dr.Ebru Balta
Ağrı İbrahim Çeçen Üniversitesi - Türkiye
ORCID: 0000-0002-2173-7189
ebrubalta2@gmail.com

Dr.Öğr. Üyesi. Arzu Uçar
Hakkari Üniversitesi - Türkiye
ORCID: 0000-0002-0099-1348
arzukapcik@gmail.com

Özet

Bu araştırmada, bilgisayar ortamında bireye uyarlanmış test (BBT) uygulamalarında, ölçme kesinliği ve test uzunluğunun, farklı test durdurma kurallarına göre değişiminin teste başlama kuralına, madde seçme ve yetenek kestirim yöntemlerine göre incelenmesi amaçlanmıştır. Araştırma, Monte Carlo simülasyon çalışması olarak gerçekleştirilmiştir. Araştırmanın amacı doğrultusunda, tepki olasılıklarının üç parametreli lojistik (3PL) model ile modellendiği 500 madde üretilmiştir. Araştırmada, teste başlama kuralı (θ=0,-1<θ< 1), madde seçim yöntemi (Maksimum Fisher Bilgisi (MFB), Kullbak-Leibler Bilgisi (KLB)) , yetenek kestirim yöntemi (Maksimum Olabilirlik Kestirimi (MOK), Beklenen Sonsal Dağılım (BSD) ve Maksimum Sonsal Dağılım (MSD)) ve testi durdurma kuralı (sabit uzunluklu (15,20), yetenek kestiriminin standart hatası (SH<.30, SH<.50)) olmak üzere her koşul için 25 yineleme ile toplam 48 (2x2x3x4) koşul incelenmiştir. Araştırma kapsamında ölçme kesinliğini belirlemede hata göstergeleri olan RMSE, yanlılık, uyum değerleri incelenmiştir. Veri üretiminde ve analizinde R yazılımı kullanılmıştır. Çalışmanın sonucunda, teste başlama kuralının koşullara göre ölçme kesinliği ve test uzunluğu açısından farklılık oluşturmadığı görülmüştür. RMSE ve yanlılık değerlerinin daha düşük elde edildiği durdurma kuralının 0,30 SH durdurma kuralı olduğu sonucuna ulaşılmıştır. BSD yetenek kestirim yönteminde MOK'a kıyasla daha düşük RMSE ve yanlılık değerleri elde edilmiştir. KLB madde seçim yönteminin MFB'ye kıyasla daha düşük RMSE ve yanlılık değerlerine sahip olduğu sonucuna ulaşılmıştır. Araştırmaya benzer bir çalışma farklı madde havuzu büyüklükleriyle gerçekleştirilebilir. Ayrıca madde havuzunun özellikleri değiştirilerek durdurma kurallarının karşılaştırılması yapılabilir. Çalışmada maddelerin kullanım sıklıkları göz önünde bulundurulmamıştır. Maddelerin kullanım sıklıklarını dikkate alan benzer çalışmalar gerçekleştirilebilir.

Anahtar Kelimeler: *Bilgisayar ortamında bireye uyarlanmış test, Madde tepki kuramı, Ölçme kesinliği, RMSE*

E-Uluslararası Eğitim Araştırmaları Dergisi

Cilt: 13, No: 1, ss. 51-68

Araştırma Makalesi

Gönderim: 2021-11-13
Kabul: 2022-01-26

Önerilen Atıf

Balta, E., & Uçar, A. (2022). Bilgisayar ortamında bireye uyarlanmış test uygulamalarında ölçme kesinliğinin ve test uzunluğunun farklı koşullar altında incelenmesi, *E-Uluslararası Eğitim Araştırmaları Dergisi,* 13(1), 51-68, DOI: https://doi.org/10.19160/e-ijer.1023098

[1] Bu araştırma makalesi 21-23 Haziran 2019 tarih aralığında Viyana, Avusturya'da düzenlenen International Conference on Research in Teaching and Education'de sözlü bildiri olarak sunulmuştur.

GİRİŞ

Eğitim ve psikoloji alanında kullanılan testlerden geçerli ve güvenilir ölçme puanları elde etmek için test geliştirme aşamasında madde ve test istatistiklerini kestirmede Klasik Test Kuramı (KTK) ve Madde Tepki Kuramı (MTK) geliştirilmiştir (Hambleton & Swaminathan, 1985; Lord, 1980). MTK, bireylerin örtük özellikleri ile maddelere verdikleri tepkiler arasındaki bağıntının matematiksel olarak gösterilmesine imkan sunmasından dolayı madde havuzu oluşturma, bireye uyarlanmış test geliştirme, test eşitleme, madde yanlılığının belirlenmesi ve bireye uyarlanmış test geliştirme konularında çözümler üretmektedir (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Thissen & Steinberg, 2009). MTK modellerinin uygulamalarından olan "bilgisayar ortamında bireye uyarlanmış test (BBT)" geleneksel kağıt kalem test uygulamalarından farklı olarak bilgisayar ortamında bireye uygulanabilen bireye uyarlanmış ölçme aracıdır. MTK modelleri, bireyin yeteneğini bireye uygulanan madde örnekleminden, madde parametrelerini de bireylerin yetenek düzeyinden bağımsız olarak kestirmesinden yani değişmez madde ve yetenek parametrelerinin elde edilmesini sağlamasından dolayı BBT uygulamalarında kullanılmaktadır. BBT uygulamaları, geleneksel kâğıt-kalem test uygulamalarından farklı olarak bilgisayar aracılığıyla gerçekleştirilmekle birlikte geleneksel uygulamalara göre MTK'nın değişmezlik özelliğini kullanarak daha yüksek ölçme kesinliği, daha düşük test süresi ve esnek uygulamalar sağlayarak her birey için madde seçimi ve yetenek kestiriminde algoritma oluşturmaktadır (Embretson & Reise, 2000; Segall,1996). BBT uygulamalarında, testi alan her birey için birey tepkilerine bağlı olarak kestirilen yetenek düzeyine uygun maddeler seçilerek farklı testler oluşturulmaktadır. Böylelikle, psikometri alanında bireysel farklılıkların ele alınması ile birlikte gündeme gelen bireyselleştirilmiş bilgisayarlı test uygulamalarında her birey kendisine optimize edilmiş farklı test almaktadır (Eggen, 2004). BBT yönteminin algoritması, seçilen maddelerin bireye sunulması ve birey tepkileri aracılığıyla yetenek düzeyinin kestirilmesini sağlayan yinelemeli bir süreçten oluşmaktadır (Wainer, 2000; Weiss, 1983).

BBT uygulamaları; madde havuzunun oluşturulması, teste başlama süreci, yetenek kestirim yöntemleri, madde seçim yöntemleri ve testi durdurma kuralları unsurlarından oluşmaktadır (Kingsbury & Zara, 1989; Orcutt, 2002; Wainer, 2000). BBT yöntemi algoritmasında öncelikle bireyin yetenek düzeyi kestirilmeye çalışılmaktadır. Böylelikle, konu alanlarına ve zorluk düzeylerine göre gruplandırılmış, madde bilgi fonksiyonları daha önce belirlenmiş ve bireylerin yetenek düzeyi (θ)'nin bütün aralıklarında bilgi sunan maddelerden oluşan geniş bir madde havuzu oluşturulup bu maddelerden birey hakkında en iyi bilgiyi verecek madde seçilerek teste başlanır (Weiss, 2004). Madde havuzunda yer alacak madde sayısına ilişkin olarak madde kullanım sıklığı (item exposure) gözönünde bulundurularak test güvenliğini zedeleyen durumların oluşmasını engellemek için geniş bir madde havuzu oluşturularak madde havuzuna sürekli olarak yeni maddeleri eklemek (pretesting) gerekmektedir (Embretson & Reise, 2000; Glas & van der Linden, 2003; Magis & Raîche, 2012; McLeod & Schnipke, 1996). Her bir yetenek düzeyine uygun güçlük düzeyinde ve ayırt edicilik düzeyi yüksek maddelerden oluşan BBT uygulamaları daha iyi sonuçlar vermektedir (Veldkamp & van der Linden, 2010; Weiss, 2004).

Testin başlangıcında seçilen ilk maddenin güçlük düzeyinin belirlenmesinde, bireyin yeteneği hakkında önceden sahip olunan bilgiler kullanılabileceği gibi bireyin yeteneği hakkında ön bilginin olmadığı durumlarda orta güçlük düzeyinde madde kullanımının psikometrik olarak daha etkin olacağı belirtilmektedir (Mills & Stocking, 1996; Sereci, 2003). İlk maddenin seçiminin ardından bireyin yeteneğini kestirmede Maksimum Olabilirlik Kestirim yöntemi (MOK; Maximum Likelihood Estimation: MLE; Birnbaum, 1968), Ağırlıklandırılmış Olabilirlik Kestirim yöntemi (AOK; Weighted Likelihood Estimation: WLE; Warm, 1989), Marjinal Maksimum Olabilirlik Kestirim yöntemi (MMOK; Marginal Maximum Likelihood Estimation: MMLE; Bock & Aitkin, 1981) ve Bayes temelli yetenek kestirim yöntemlerinden Beklenen Sonsal Dağılım yöntemi (BSD; Expected a Posteriori: EAP; Bock & Aitkin, 1981), Maksimum Sonsal Dağılım yöntemi (MSD; Maximum a Posteriori: MAP; Samejima, 1977) sıklıkla kullanılmaktadır (Baker & Kim, 2004; Embretson & Reise, 2000). MOK yöntemi, olabilirlik fonksiyonu kullanılarak birey hakkında en fazla bilgiyi veren maddeyi seçmeye dayanan yetenek kestirim yöntemidir. Bireylerin maddelere verdiği tepkilerin birbirinden bağımsız olduğu durumda tepki olasılıklarının çarpımı olabilirlik fonksiyonu olarak tanımlanmakta ve bu fonksiyonu en yüksek yapan yetenek düzeyi belirlenerek en çok olabilirlik kestirimi yapılmaktadır. Eşitlik 1'de bireyin tepki örüntüsüne bağlı olarak koşullu olabilirlik fonksiyonu gösterilmektedir.

$$L(u_{s1}, u_{s2}, \dots, u_{sl} \mid \theta_s) = \prod_{i=1}^{l} P_i(\theta_s)^{u_{si}} Q_i(\theta_s)^{1-u_{si}} \qquad i = 1,2,3,\dots,l \tag{1}$$

Eşitlik 1'de; $P(\theta)$, bireyin belli bir θ düzeyinde maddeye doğru tepki verme olasılığını ve $Q(\theta)$, bireyin belli bir θ düzeyinde maddeye yanlış tepki verme olasılığını ifade etmektedir. MOK yöntemi ile yetenek kestirimi yapabilmek için en az bir doğru ve bir yanlış tepkiden oluşan birey tepki örüntüsüne ihtiyaç duyulmaktadır. Böylelikle, MOK yöntemi, tümü ile doğru tepki örüntüsü olduğu durumda pozitif sonsuzda monoton artan, tümü ile yanlış tepki örüntüsü olduğu durumda negatif sonsuzda monoton azalan bir fonksiyon gösterdiği için fonksiyonu en yüksek yapan değeri bulmanın mümkün olamamasından dolayı yetenek kestirimi yapamamaktadır. Birey için yetenek kestirimi yapılırken madde karakteristik eğrilerinin sıfır değerini içermesinden dolayı madde karakteristik eğrilerinin doğal logaritmalarının alınıp toplanmasıyla elde edilen Eşitlik 2'de gösterilen log-L fonksiyonunu en büyük yapan değer alınmaktadır.

$$-\log L(u_{s1}, u_{s2}, \dots, u_{sl} \mid \theta_s) = \sum_{i=1}^{l} u_{si} \log[P_i(\theta)] + (1 - u_{si}) \log[Q_i(\theta)] \tag{2}$$

Eşitlik 2'de; i madde indeksini, l bireyin cevapladığı madde sayısını göstermektedir. Madde sayısının fazla olduğu ve geniş örneklem büyüklüğüne sahip verilerde Newton-Raphson iterasyon yöntemi kullanılarak bireyin yetenek kestirimi yapılmaktadır (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991; Wang & Vispoel,1998). MOK yetenek kestiriminin sınırlıklarını gidermek için önsel bir yetenek dağılımı kullanılarak bayesci yöntemler geliştirilmiştir (Baker & Kim, 2004; Hambleton & Swaminathan, 1985; Lord & Stocking, 1988). Bayesci yöntemler, testi alan her birey için yetenek dağılımlarını ortalaması 0, standart sapması 1 olan normal dağılımda olduğunu varsayarak önsel dağılımı birey ilk maddeyi cevapladıktan sonra birey tepki örüntüsünden elde edilen olabilirlik bilgisi ile birleştirerek sonsal yetenek dağılımını oluşturur. Test sonlanıncaya kadar her maddenin cevaplanmasından sonra elde edilen sonsal dağılım bir sonraki madde için önsel dağılım olarak kullanılır (Wang & Vispoel, 1998). Bayesci yöntemlerde önsel ve sonsal dağılımların karakteristiklerine göre farklı yöntemler geliştirilmiştir. MSD yöntemi, birey yetenek düzeyi kestirimini önselin ortalamasına doğru çeken ve madde sayısının az olduğu durumlarda kestirimin standart hatasını düşüren önsel dağılımlar kullanmaktadır. Böylelikle, yanlış önsel dağılım kullanımında birey yetenek düzeyine ilişkin yanıltıcı sonuçlar üretebilmektedir (Embretson & Reise, 2000). BSD yetenek kestirim yöntemi, yetenek kestriminde sonsal dağılımın ortalamasından faydalanarak yetenek düzeyi kestirimi sağlar. BSD yetenek kestirimi Eşitlik 3 ile ifade edilmektedir.

$$EAP(\theta) = E(\theta|u) = \int_{-\infty}^{\infty} P(\theta|u)\, \theta\, d\theta \tag{3}$$

Eşitlik 3'de; $P(\theta|u)$,sonsal dağılımı ,$E(\theta|u)$, sonsal dağılımın ortalamasını, $\int_{-\infty}^{\infty} \theta\, d\theta$ ise bütün yetenek değerlerinin üzerindeki alanı ifade etmektedir. Eşitlik 3 incelendiğinde, BSD yetenek kestirimi yönteminin MOK ve MSD yetenek kestirim yöntemlerinin aksine karışık ve iteratif bir süreç içermediği ve tüm yetenek düzeyleri için sonlu bir kestirim yaptığı ve böylece birey tepilerinin tümü doğru ya da yanlış olduğu durumda da yetenek kestirimi yaptığı görülmektedir (Bock & Aitkin, 1981; Embretson & Reise, 2000; Hambleton vd., 1991; Ho, 2010). BBT uygulamalarında yetenek kestirim yöntemlerinden hangisinin daha iyi olduğuna karar vermede madde seçim yöntemleri, madde seçiminde içerik dengesi, madde kullanım sıklığının belirlenmesi gibi değişkenler önemli rol oynamaktadır.

BBT uygulamalarında, bireylerin ilk yetenek düzeyleri kestirildikten sonraki aşama testin devamında bireye uygulanacak maddelerin seçilmesidir. BBT için en basit madde seçim algoritması random madde seçimidir. Random madde seçimi, maddeler ve bireyler hakkında bilgi kullanılmadan gerçekleştirildiği için etkili bir yöntem olarak görülmemektedir. Daha etkili bir yaklaşım olarak kesme puanı temelli ve kestirim temelli olmak üzere zeki madde seçim yaklaşımı geliştirilmiştir. Klasik Test Kuramı (KTK) temelli maksimum ayırıcılık, bilgi kazanımı ve minumum beklenen değer, Madde Tepki Kuramı (MTK) temelli Maksimum Fisher Bilgisi (MFB; Maximum Fisher Information: MFI), Kullback Leibler Bilgisi (KLB; Kullback-Leibler Information: KLI) ve log-odds ratio kesme puanı temelli yöntemlerde kesme

noktasında en yüksek bilgiyi veren maddeler seçilmektedir (Thompson, 2007b). MFB, her maddeye verilen tepkiden sonra yapılan yetenek kestirimi (interim yetenek kestirimi) $\hat{\theta}$ 'da daha önce uygulanan m-1 tane madde için $I[\hat{\theta}_{m-1}]$ için en yüksek değeri veren maddeyi bulmayı amaçlayarak, tek bir noktada (kestirilen geçici yetenek düzeyinde) bilginin maksimize edilmesini sağlar (Embretson & Reise, 2000). Eşitlik 4'de üç parametreli lojistik model (3PLM) için MFB'ye dayalı madde seçimi gösterilmektedir.

$$I_i[\hat{\theta}_{m-1}] = \frac{(Da_i)^2(1-c_i)}{\left[c_i + e^{Da_i(\hat{\theta}_{m-1}-b_i)}\right]\left[1+e^{-Da_i(\hat{\theta}_{m-1}-b_i)}\right]^2} \quad (4)$$

Eşitlik 4 incelendiğinde, Fisher bilgisinin, belirli bir yetenek düzeyi civarındaki bilginin ölçüsünü (yerel bilgi) kullandığı ve madde ayırt edicilik düzeyi (a_i) arttıkça sağladığı bilgi düzeyinin de arttığı görülmektedir Böylelikle, MFB madde seçme yönteminin, birey için maksimum test bilgisi sağlamaya yönelik yüksek düzeyde ayırt edici maddeleri seçmesinin madde havuzunun yanlı kullanılmasına ve test uygulamasının başlangıcında az sayıda madde uygulanmasından kaynaklı olarak interim yetenek kestirimini yanlı belirlemesine yol açabilmektedir (Han, 2009; Ho, 2010).

KLB, MTK'da madde seçme amacı ile global bilginin ölçüsü olarak (Chang & Ying, 1996), çok boyutlu MTK'ya dayalı test geliştirmede (Veldkamp ve van der Linden, 2002) ve bilişsel tanıya dayalı değerlendirmelerde (Tatsuoka & Ferguson, 2003) madde seçim algoritması olarak kullanılmıştır. KLB, global bilgiyi kullanarak yetenek düzeylerinin geniş ranjı boyunca farklılaşma gücünü gösterir. Eşitlik 5'de, BBT uygulamalarına uyarlanan i maddesi için KLB ifade edilmektedir.

$$K_i(\theta \| \theta_0) = P_i(\theta_0) \log\left[\frac{P_i\ (\theta_0)}{P_i\ (\theta)}\right] + [1 - P_i\ (\theta_0)] \log\left[\frac{1-P_i\ (\theta_0)}{1-P_i\ (\theta)}\right] \quad (5)$$

Eşitlik 5'de; θ_0 ,gerçek yetenek düzeyini ifade etmektedir. Eşitlik 5 incelendiğinde, KLB ile iki yetenek düzeyi (θ, θ_0) arasında bir maddenin değişme kapasitesinin karakterize edildiği görülmektedir. Yetenek düzeyinin konumu hakkında yeterli bilgiye sahip olunmadığı durumlarda, KLB'nin global bilgiyi kullanmasından kaynaklı olarak BBT uygulamalarının başlangıç aşamasında kullanımı önerilmektedir. KLB'nin kullanımı, kısa test uzunluklarında ve testin başlangıcında yetenek kestirim hata düzeylerini düşürmektedir (Han, 2009; Ho, 2010). KLB ve MFB kesme puanı yerine bireyin kestirilen geçici yetenek düzeyini kullanan kestirim temelli yaklaşımlarda da kullanılabilmektedir (Eggen, 1999).

BBT uygulamalarında, yetenek kestirimlerinin yanlılığı ve testin standart hatası üzerinde farklı etkiler oluşturan, testin amacına ve madde havuzunun karakteristik özelliğine göre farklılaşan sabit ve değişken test uzunluğu olmak üzere iki farklı testi durdurma kuralı bulunmaktadır (Babcock & Weiss, 2012; Blais & Raiche, 2010; Segall, 2004; Sereci, 2003; Weiss & Kingsbury, 1984). Sabit test uzunluklu test durdurma kuralında, ölçme kesinliğinin derecesi ve zaman kaybı göz ardı edilerek testin daha yüksek kapsam geçerliğine sahip olması için önceden belirlenen madde sayısına ulaşıldığında test sonlandırılmaktadır. Küçük ölçekli BBT uygulamalarında madde havuzunda yer alan madde sayısının az olduğu durumlarda sabit test uzunluklu durdurma kuralı tercih edilmektedir (Babcock & Weiss, 2009). Standart Hata (SH), Teta Değişimi ve Minimum Bilgi (MB) değişken test uzunluğu durdurma kuralında bireylerin tepki örüntülerine göre farklı sayıda madde ile test sonlandırılarak ölçmenin etkililiği (madde havuzunun etkili kullanılıp birey yeteneğinin göreli olarak daha az madde ile kestirilmesi) sağlanabilmektedir (Babcock & Weiss, 2009; Weiss & Kingsbury, 1984). SH durdurma kuralında, bireyin yetenek kestirimi belli bir kesinlik düzeyine ulaşıncaya kadar kabul edilebilir standart hata için bilgi değeri yüksek madde uygulanmaktadır (Hambleton vd., 1991; Wang & Wang, 2001; Weiss & Kingsbury, 1984). MB durdurma kuralı, Fisher'ın madde bilgi fonksiyonuna dayanan kestirilen yetenek seviyesi ile ilgili madde havuzunda iyi bilgi verecek madde kalmadığı durumda testin sonlandırılmasında kullanılmaktadır (Babcock & Weiss, 2009). Theta değişimi durdurma kuralında ise bireyin daha önceden belirlenen yetenek düzeyi ve bu yetenek düzeyine yakınsayan yetenek düzeyi değeri test durdurma kuralı olarak belirlenir (Babcock & Weiss, 2012; Weiss & Kingsbury,1984).

Yurtdışı ve Türkiye genelinde geniş ölçekli BBT uygulamalarının (GMAT (Graduate Management Admission Test), GRE (Graduate Record Examination), TOEFL (Test of English as a Foreign Language) vb.) hızla artış gösterdiği görülmektedir. Ayrıca, küresel çapta görülen COVID 19 pandemi dönemi ile birlikte bireylerin hayatlarında önemli kararların alınmasında önemli yer tutan, geniş ölçekli ve yüksek riskli sınavların uygulanması ve uzaktan eğitim sürecinde, ölçme ve değerlendirme uygulamalarında kullanılan

çevrim içi sınav uygulamalarının gerekliliği günümüzde giderek hız kazanmaktadır. BBT'a yönelik çalışmalarda, madde seçim algoritmalarının, teste başlama ve sonlandırma kurallarının, birey yeteneklerini kestirim yöntemlerinin, madde havuzu genişliğinin, madde kullanım sıklığı kontrol yöntemlerinin ölçmenin kesinliği ve etkilerinin incelenildiği görülmektedir. İlgili literatür incelendiğinde, BBT uygulamalarında, madde havuzu özelliklerinin, teste başlama kurallarının, yetenek kestirim ve madde seçim yöntemlerinin ve testi durdurma kurallarının birlikte ölçme kesinliği ve etkililiği üzerine etkisini inceleyen çalışmalarda (Babcock & Weiss, 2012; Blais & Raiche, 2002; Chang & Ansley, 2003; Choi, Grady & Dodd, 2010; Eroğlu & Kelecioğlu, 2015; Kalender, 2011; Ivie, 2007; İşeri 2002; Simms & Clark, 2005; Yi, Wang & Ban, 2001) bazı araştırmacıların (Chang & Ansley, 2003; Eroğlu & Kelecioğlu, 2015; Yi, Wang & Ban, 2001), kağıt-kalem test uygulamaları ile benzerlik gösteren sabit uzunluklu test durdurma kuralının değişken uzunluklu test durdurma kuralına göre yanlı sonuçlar gösterdiği sonucuna ulaştığı görülmektedir. Bununla birlikte bazı araştırmalarda (Baker & Kim, 2004; Bock & Aitkin, 1981; Bock & Mislevy, 1982; Embretson & Reise, 2000; Eroğlu & Kelecioğlu, 2015; Hambleton & Swaminathan, 1985, 1991; Ho, 2010; Lord,1983; Lord & Stocking, 1988; Warm, 1989), ilk maddenin seçiminin ardından birey yeteneğini kestirmede kullanılan bayesci yöntemlerin olabilirliğe dayalı yöntemlere göre daha kesin yeterlik tahmini yaptığı belirtilmiş olmakla birlikte yetenek kestirim yöntemlerinden hangisinin daha iyi olduğuna karar vermede madde seçim yöntemleri, madde seçiminde içerik dengesi, madde kullanım sıklığının belirlenmesi değişkenler gibi önemli rol oynadığı belirtilmiştir. Bununla birlikte, yapılan araştırmalar (Deng, Ansley & Chang, 2010; Han, 2010; Kalender, 2011;Şahin & Özbaşı,2017; İşeri, 2002; Wen, Chang & Hau, 2000; Yi & Chang, 2003) incelendiğinde, madde seçim yöntemlerinin üstünlüklerinin belirlenmesinin önem teşkil ettiği görülmektedir. Böylelikle, BBT uygulamalarında, ölçme kesinliği ve test uzunluğunun farklı faktörlere göre değişimini incelemenin ölçme ve değerlendirme alanyazınına katkı sunacağı düşünülmektedir. Bu araştırmanın amacı, bilgisayar ortamında bireye uyarlanmış test uygulamalarında, ölçme kesinliği ve test uzunluğunun, farklı test durdurma kurallarına göre değişiminin teste başlama kuralına, madde seçme ve yetenek kestirim yöntemlerine göre incelenmesidir. Bu amaç doğrultusunda aşağıdaki sorulara cevap aranmıştır:

BBT uygulamalarında, ölçme kesinliği ve test uzunluğu, farklı durdurma kurallarına göre;

1- Başlangıç kuralı olarak θ=0 ve -1< θ <1 kullanıldığı durumda nasıl değişmektedir?
2- Madde seçim yöntemi olarak Maksimum Fisher Bilgisi (MFB) ve Kullback Leibler Bilgisi (KLB) kullanıldığı durumda nasıl değişmektedir?
3- Yetenek kestirim yöntemi olarak Maksimum Olabilirlik Kestirim yöntemi (MOK), Maksimum Sonsal Dağılım yöntemi (MSD) ve Beklenen Sonsal Dağılım yöntemi (BSD) kullanıldığı durumda nasıl değişmektedir?

YÖNTEM

Bu araştırmada, simülasyon verileri kullanılarak bilgisayar ortamında bireye uyarlanmış test (BBT) uygulamalarında, ölçme kesinliği ve test uzunluğunun farklı test durdurma kurallarına göre değişimi; teste başlama kuralına, madde seçme ve yetenek kestirim yöntemlerine göre incelemek için Monte-Carlo simülasyon çalışması gerçekleştirilmiştir. Araştırmada, gerçek veri ile çalışmada ele alınan koşulların tümünün sağlanmasının mümkün olmamasından dolayı simülasyon verisi kullanılmıştır. Bu çerçevede mevcut araştırma bir Monte-Carlo simülasyon çalışması olarak değerlendirilebilir (Harwell, Stone, Hsu & Kirisci, 1996).

Verilerin Üretilmesi

Araştırmada, BBT uygulamalarında, ölçme kesinliği ve test uzunluğunun farklı test durdurma kurallarına göre değişimini teste başlama kuralına, madde seçme ve yetenek kestirim yöntemlerine göre incelemek üzere Monte-Carlo simülasyon çalışması yürütülmüştür. Literatürde de BBT uygulamaları çalışmalarında sıklıkla simülatif veriye başvurulduğu görülmektedir (Bulut & Kan, 2012; Evans, 2010; Kalender, 2011; McDonald, 2002; Scullard, 2007). BBT veri üretimi, R yazılımında *"catR"* paketi kullanılarak gerçekleştirilmiştir. BBT'de etkili bir yetenek tahmininin yapılabilmesi için en az 100 maddeden oluşan madde havuzunun olması gerektiği literatürde ifade edilmesiyle birlikte, madde havuzu büyüklüğünün

test uzunluğunun en az altı ile on iki kat fazlası madde içermesi gerektiği belirtilmiştir (Stocking, 1992; Urry, 1977) Bu araştırmanın madde havuzu, tepki olasılıklarının üç parametreli lojistik (3PL) model ile modellendiği 500 maddeden oluşmaktadır.

Araştırmada, madde seçme yöntemi olarak, Maksimum Fisher Bilgisi (MFB) ve Kullbak-Leibler Bilgisi (KLB) yöntemleri seçilmiştir. BBT uygulamalarında, yüksek ayırt edicilik düzeyindeki maddeleri seçmeye eğilimli olan MFB madde seçme yönteminin testin başlangıcında yetenek kestiriminde yetersiz kaldığı belirtilmektedir (Deng, Ashley & Chang, 2010; Chang & Ying, 1999; Han, 2009; Veldkamp, 2012; Weissman, 2003). Böylelikle, KLB madde seçme yönteminin MFB yöntemine göre daha kesin yeterlik kestirimleri sağladığı sonucuna ulaşan araştırmalar (Eggen, 1999; Han, 2009; Reckase, 1983; Spray & Reckase, 1994; Sulak & Kelecioğlu, 2019) dikkate alınarak yöntemler karşılaştırılmalı olarak ele alınmıştır. Araştırmada, yetenek kestirim yöntemi olarak Maksimum Olabilirlik Kestirimi (MOK), Beklenen Sonsal Dağılım (BSD) ve Maksimum Sonsal Dağılım (MSD) yöntemleri ele alınmıştır. BBT uygulama sürecinde yetenek kestirim yöntemlerinin madde seçim yöntemlerini etkilediği belirtilmektedir (Bock & Mislevy, 1982; Wang & Visposel, 1998; Warm, 1989). MOK yönteminin bayesci yöntemlere kıyasla yanlı yetenek kestirimi yaptığını belirten çalışmalar (Bock & Mislevy, 1982; Lord,1983; Warm, 1989) dikkate alınarak MOK yöntemi ve bayesci yöntemlerden BSD ve MSD yöntemlerinin madde seçme yöntemleri ile birlikte test durdurma kuralına göre ölçme kesinliğini ve test uzunluğunu nasıl etkilediği incelenmiştir.BBT uygulamalarında testi başlatma aşaması; bireyin yeteneği hakkındaki ön bilgiye göre, kolay maddelerle ya da orta güçlükte maddelerle başlama şeklinde farklı yaklaşımlarla gerçekleşebilir (Hambleton & Xing, 2006; Parshall, Spray, Kalohn & Davey, 2002; Thompson & Weiss, 2011). Bu çalışmada ise orta güçlükte maddelerle başlamak için θ=0, kolay ve zor maddelerle başlamak için ise -1<θ<1 koşulları dikkatte alınmıştır. Testi durdurmada, sabit uzunluk, yetenek düzeyinin standart hatası kuralları kullanılmıştır. Sabit uzunluk test durdurma kuralı için, Blaise ve Raiche (2002) çalışmasında en az 13 madde uygulanmasını önermekle birlikte Weissman (2003) çalışmasında, 5, 10, 15 ve 25 madde, Wen, Chang ve Hau (2010), 15 madde, Han (2010), 10, 20 ve 40 madde, Sulak (2013), 5, 10,20 30 ve 40 madde, Eroğlu ve Kelecioğlu (2015), 15 ve 20 madde olarak ele almıştır. Böylelikle, bu çalışmada, test durdurma kuralı olarak 15 ve 20 madde olacak şekilde farklı iki koşul belirlenmiştir. Kestirilen yetenek düzeyinin standart hatası durdurma kuralı için madde havuzunda tüm yeterlik düzeyine ilişkin madde bulunduğunda standart hata değerinin belirlenmesinde güvenirlik değerininin karesinin göz önünde bulundurulmasıyla birlikte standart hatanın 0,40' a eşit veya daha düşük olduğu durumlarda ölçme kesinliğinin daha yüksek olduğu belirtilmiştir (Babcock & Weiss, 2012; Blaise & Raiche, 2002; Wang, Hanson & Lau, 1999). Böylelikle, bu araştırmada, test durdurma kuralı olarak SH<0.50, SH<0.30 olduğu BBT koşulları ele alınmıştır. Araştırmada belirlenen değişkenler ve düzeyleri Çizelge 1'de yer almaktadır.

Çizelge 1. *Simülasyon Deseni*

	Faktörler (Değişkenler)	Koşullar (Düzeyler)	Koşul Sayısı
Sabit Faktörler	Örneklem Büyüklüğü	1000	1
	Madde Havuzu Büyüklüğü	500	1
Değişimlenen Faktörler	Başlama Kuralı	Θ=0, -1< Θ <1	2
	Madde Seçme Yöntemi	KLB, MFB	2
	Yetenek Kestirim Yöntemi	MOK, MSD, BSD	3
	Durdurma Kuralı	SH<0.50, SH<0.30, Sabit uzunluk=15, Sabit uzunluk=20	4

Çalışmada, teste başlama kuralı (2 düzey), madde seçim yöntemi (2 düzey) , yetenek kestirim yöntemi (3 düzey) ve testi durdurma kuralı (4 düzey) olmak üzere toplam 48 (2x2x3x4) koşul incelenmiştir. Araştırmanın genelinde 48 koşul için R. 3.4.0 programı çalıştırılmıştır. Örneklem yanlılığının ortadan kaldırılmasına yönelik (Harwell vd.,1996) araştırmada her koşul için 25 yineleme gerçekleştirilmiştir.

Verilerin Analizi

Verilerin analizi, bağımlı değişkenler olan test uzunluğu ve ölçme kesinliğine ilişkin değerler (RMSE, yanlılık ve uyum) her koşul için 25 yinelemenin ortalaması olacak şekilde araştırmacılar tarafından yazılan fonksiyonlarla R yazılımı kullanılarak gerçekleştirilmiştir. Araştırma kapsamında; ölçme kesinliğinin belirlenmesinde, hata göstergelerinden RMSE, yanlılık, uyum değerleri incelenmiştir. Aşağıda hata değerlerine ilişkin kısaca bilgi verilmiştir.

RMSE (Root Mean Squared Error): RMSE, BBT simülasyonu sonucunda bireyin kestirilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki mutlak farka ilişkin istatistik olup tüm koşullar için hesaplanan hataların karesinin ortalamasının karekökünü göstermektedir. RMSE değeri aşağıdaki eşitlik yardımıyla hesaplanır.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}} \qquad (6)$$

Eşitlik 6'da, n, toplam birey sayısını; θ_i, i. bireyin gerçek yetenek düzeyi değeri; $\hat{\theta}_i$, i. bireyin kestirilen yetenek düzeyi değerini göstermektedir.

Yanlılık (Bias): BBT simülasyonu sonucu**,** bireyin kestirilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki ortalama anlamlı farklılık istatistiğidir. RMSE değerinin, Yanlılık değeri ile standart hata değerini içerdiği görülmektedir($RMSE^2 = Bias^2 + SE^2$)

Uyum (Fidelity): Uyum katsayısı, bireylerin gerçek yetenek düzeyi ile kestirilen yetenek düzeyi arasındaki Pearson korelasyon katsayısı olarak tanımlanır (Weiss,1982). Bireye ilişkin kestirilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki korelasyon katsayısı Eşitlik 7 ile gösterilmiştir.

$$r = \frac{cov(\hat{\theta}, \theta)}{SS(\hat{\theta})SS(\theta)} \qquad (7)$$

Eşitlik 7'de, *r* değerinin yüksek elde edilmesi; kestirilen yetenek düzeyi ile gerçek yetenek düzeyi arasındaki yüksek uyumu göstermektedir.

BULGULAR

Bu bölümde, Bilgisayar ortamında bireye uyarlanmış test (BBT) uygulamasında, test durdurma kuralına göre madde seçmede kullanılan yöntemlerin, başlangıç kurallarının ve yetenek kestirim yöntemlerinin ölçme kesinliği ve test uzunluğu açısından değerlendirmelerine ilişkin bulgulara yer verilmiştir. Çizelge 2'de, BBT uygulamasında çeşitli faktörler altında ölçme kesinliği ve test uzunluğunun değişimine ilişkin RMSE, yanlılık, uyum katsayısı ve test uzunluğu değerlerinin ortalama değerleri gösterilmiştir.

BBT testinin başlama kuralı θ=0 olduğunda en düşük RMSE değeri, madde seçme yöntemi KLB, yetenek kestirim yöntemi MOK ve durdurma kuralı SH<0.30 seçildiğinde gözlenmiştir. Aynı başlama kuralı koşulunda en düşük yanlılık değeri ise madde seçme yöntemi KLB, yetenek kestirim yöntemi MSD ve testi durdurma kuralı SH<0.30 olarak seçildiğinde elde edilmiştir. Başlama kuralı θ=0 olduğunda en yüksek uyum katsayısı, madde seçme yöntemi KLB, yetenek kestirim yöntemi MOK seçildiğinde gözlenmiştir.

Tablo 2 incelendiğinde, başlama kuralı -1<θ<1 olduğunda, en düşük RMSE değeri madde seçim yöntemi KLB, yetenek kestirim yöntemi KLB ve testi durdurma kuralı SH<0.30 seçildiğinde elde edilmiştir. Aynı başlama koşulu altında en düşük yanlılık değeri, madde seçim yöntemi KLB, yetenek kestirim yöntemi MSD ve sonlandırma kuralı sabit uzunluk 15 olarak seçildiğinde gözlenmiştir. Başlama kuralı -1<θ<1 olduğunda en yüksek uyum katsayısı ise madde seçim yöntemi KLB, yetenek kestirim yöntemi MOK ve testi sonlandırma kuralı SH<0.30 olduğunda elde edilmiştir. Tablo 2 incelendiğinde; ele alınan tüm koşullarda, en düşük RMSE değeri, başlama kuralı θ=0, madde seçme yöntemi KLB, yetenek kestirim yöntemi MOK ve testi durdurma kuralının SH<0.30 olduğu durumda 0.296 olarak elde edilmiştir. Ayrıca, başlama kuralının, *-1<θ<1,* madde seçme yönteminin KLB, yetenek kestirim yönteminin BSD ve testi durdurma kuralının SH<0.30 olarak seçildiği durumda benzer RMSE değeri gözlenmiştir. Çalışmada, tüm koşullar altında, en yüksek RMSE değeri 0,86 ile başlama kuralının θ=0, madde seçme yönteminin MFB, yetenek kestirim yönteminin MOK ve testi durdurma kuralının 15 maddelik sabit uzunluk olduğu durumda elde edildiği görülmektedir. En düşük yanlılık değerinin ise tüm koşullar altında 0 değeri ile teste başlama kuralının θ=0, madde seçme yönteminin KLB, yetenek kestirim yöntemi MSD ve testi durdurma kuralının SH<0.30 olduğu koşulda gözlendiği görülmektedir. Tüm koşullar altında en yüksek

yanlılık değeri ise 0.056 olarak, başlama kuralının θ=0, madde seçme yönteminin MFB, yetenek kestirim yöntemi MOK ve testi durdurma kuralının SH<0.50 olduğu koşulda elde edildiği görülmektedir.

Çizelge 2. *Koşullara ilişkin RMSE, yanlılık, uyumkKatsayısı ve Test Uzunluğu Ortalama Değerleri*

Başlama Kuralı	Madde Seçim Yöntemi	Yetenek Kestirim Yöntemi	Durdurma Kuralı	RMSE	Yanlılık	Uyum Katsayısı	Test Uzunluğu
θ=0	MFB	MOK	SH<0.50	0.774	0.056**	0.789	14
			SH<0.30	0.505	0.006	0.889	40**
			Sabit uzunluk=15	0.860**	0.051	0.758	15
			Sabit uzunluk=20	0.760	0.043	0.801	20
		MSD	SH<0.50	0.749	0.051	0.721*	8*
			SH<0.30	0.490	0.001	0.880	33
			Sabit uzunluk=15	0.635	0.021	0.801	15
			Sabit uzunluk=20	0.486	0.012	0.877	20
		BSD	SH<0.50	0.494	-0.025	0.874	9
			SH<0.30	0.300	-0.004	0.955	33
			Sabit uzunluk=15	0.402	-0.012	0.918	15
			Sabit uzunluk=20	0.361	-0.012	0.935	20
	KLB	MOK	SH<0.50	0.492	0.006	0.895	13
			SH<0.30	0.296*	0.007	0.960**	38
			Sabit uzunluk=15	0.467	0.006	0.911	15
			Sabit uzunluk=20	0.409	0.007	0.929	20
		MSD	SH<0.50	0.507	-0.015	0.867	9
			SH<0.30	0.304	0.000*	0.954	33
			Sabit uzunluk=15	0.407	-0.007	0.916	15
			Sabit uzunluk=20	0.365	-0.008	0.933	20
		BSD	SH<0.50	0.490	-0.009	0.876	9
			SH<0.30	0.299	-0.007	0.956	34
			Sabit uzunluk=15	0.409	-0.012	0.915	15
			Sabit uzunluk=20	0.364	-0.012	0.934	20
-1<θ<1	MFB	MOK	SH<0.50	0.491	0.018	0.896	13
			SH<0.30	0.302	0.008	0.958	39
			Sabit uzunluk=15	0.485	0.018	0.905	15
			Sabit uzunluk=20	0.413	0.009	0.928	20
		MSD	SH<0.50	0.512	-0.004	0.864	8*
			SH<0.30	0.309	0.003	0.953	32
			Sabit uzunluk=15	0.413	-0.005	0.914	15
			Sabit uzunluk=20	0.362	-0.005	0.934	20
		BSD	SH<0.50	0.492	-0.018	0.875	9
			SH<0.30	0.298	-0.008	0.956	34
			Sabit uzunluk=15	0.403	-0.016	0.918	15
			Sabit uzunluk=20	0.367	-0.009	0.932	20
	KLB	MOK	SH<0.50	0.490	0.006	0.898	13
			SH<0.30	0.298	0.008	0.959	38
			Sabit uzunluk=15	0.463	0.015	0.911	15
			Sabit uzunluk=20	0.405	0.002	0.930	20
		MSD	SH<0.50	0.509	-0.010	0.865	9
			SH<0.30	0.303	-0.002	0.954	33
			Sabit uzunluk=15	0.414	0.001	0.913	15
			Sabit uzunluk=20	0.367	0.003	0.932	20
		BSD	SH<0.50	0.494	-0.017	0.874	9
			SH<0.30	0.296*	-0.006	0.957	34
			Sabit uzunluk=15	0.405	-0.010	0.917	15
			Sabit uzunluk=20	0.363	-0.013	0.934	20

*** en yüksek değeri; * en düşük değeri göstermektedir.*

Tablo 2'ye göre en yüksek ortalama uyum katsayısı değerinin, başlama kuralının θ=0, madde seçme yönteminin KLB, yetenek kestirim yönteminin MOK ve testi durdurma kuralının SH<0.30 olduğu koşulda, 0.96 olarak gözlendiği görülmektedir. Başlama kuralının θ=0, madde seçme yönteminin MFB, yetenek kestirim yöntemi MSD ve testi durdurma kuralının SH<0.50 durumda ise en düşük ortalama uyum katsayısı 0.721 olarak elde edilmiştir. Koşullara ilişkin ortalama test uzunlukları incelendiğinde, testin en az 8 madde ile, başlama kuralının θ=0, madde seçme yönteminin MFB, yetenek kestirim yöntemi MSD ve testi durdurma kuralının SH<0.50 olduğu ve başlama kuralının -1<θ<1, madde seçme yönteminin MFB, yetenek kestirim yöntemi MSD ve testi durdurma kuralının SH<0.50 koşullarda sonlandırıldığı

gözlemlenmiştir. Başlama kuralının θ=0, madde seçme yönteminin MFB, yetenek kestirim yönteminin MOK ve testi durdurma kuralının SH<0.30 olduğu durumda ise testin en çok (40) madde ile sonlandırıldığı görülmektedir.

TARTIŞMA, SONUÇ VE ÖNERİLER

Bu araştırmada, bilgisayar ortamında bireye uyarlanmış test (BBT) uygulamalarında, ölçme kesinliği ve test uzunluğunun farklı test durdurma kurallarına göre değişimi, teste başlama kuralına, madde seçme ve yetenek kestirim yöntemleri çerçevesinde incelenmiştir. Sonuçlar, durdurma kuralları çerçevesinde verilmiştir.

Teste Başlama Kuralının θ=0 veya -1<θ<1 Belirlenmesine İlişkin Sonuçlar

Teste başlama kurallarının her ikisinde de, madde seçim yöntemi MFB ve yetenek kestirim yöntemi MOK seçildiğinde; en düşük RMSE değerlerinin sabit uzunluk durdurma kurallarında ve 0.30 SH durdurma kuralında elde edildiği görülmektedir. Yanlılık değerleri ve testte yer alan madde sayıları arasında koşullara ilişkin önemli farklılık görülmemektedir.

Madde seçim yöntemi MFB, yetenek kestirim yöntemi MSD olduğu durumda, her iki teste başlama koşulunda benzer RMSE değerlerinin sabit uzunluk durdurma kuralında ve en düşük RMSE değerinin ise 0.30 SH durdurma kuralında elde edildiği görülmektedir. Elde edilen bu sonuç Eroğlu ve Kelecioğlu (2015) çalışmasında testin orta güçlükte madde (θ=0) başladığı ve en düşük SH değerinde en düşük RMSE elde edildiği sonucuyla tutarlılık göstermektedir. Yanlılık değerleri karşılaştırıldığında, θ=0 teste başlama kuralı için 0.50 SH durdurma kuralında elde edilen yanlılık değerinin, -1<θ<1 teste başlama kuralında elde edilen yanlılık değerinden daha büyük değere sahip olduğu görülmektedir. Testte yer alan madde sayılarına ilişkin önemli bir farklılık görülmemektedir.

Madde seçim yöntemi MFB ve yetenek kestirim yöntemi BSD seçildiği koşullar için en düşük RMSE değerlerinin, her iki teste başlama kuralında da 0.30 SH durdurma kuralında elde edildiği görülmektedir. Koşullara ilişkin yanlılık değerleri ve testte yer alan madde sayılarının önemli bir farklılık göstermediği görülmektedir. Madde seçim yöntemi KLB ve yetenek kestirim yöntemi MOK seçildiğinde en düşük RMSE değerlerinin, her iki teste başlama kuralında da 0.30 SH durdurma kuralında elde edildiği görülmektedir. Yanlılık değerleri karşılaştırıldığında, θ=0 teste başlama kuralı için sabit uzunluk 15 madde ve 0.50 SH durdurma kurallarında en düşük değerlere sahip olduğu görülmektedir. -1<θ<1 teste başlama kuralı için yanlılık değerleri karşılaştırıldığında ise en düşük değerin sabit uzunluk 20 madde durdurma kuralında elde edildiği görülmektedir. Testte yer alan madde sayıları arasında farklılık görülmemektedir.

Madde seçim yöntemi KLB ve yetenek kestirim yöntemi MSD seçildiği durumlarda en düşük RMSE değerlerinin her iki teste başlama kuralında 0.30 SH durdurma kuralında elde edildiği görülmektedir. θ=0 teste başlama kuralı için; 0.30 SH durdurma kuralında yanlılık değerinin en düşük değere sahip olduğu görülmektedir. -1<θ<1 teste başlama kuralı için yanlılık değerleri karşılaştırıldığında ise en düşük yanlılık değerinin sabit uzunluk 15 madde durdurma kuralında elde edildiği görülmektedir. Testte yer alan madde sayılarının ise koşullara göre önemli bir farklılık göstermediği görülmektedir.

Madde seçim yöntemi KLB ve yetenek kestirim yöntemi BSD seçildiğinde, her iki teste başlama kuralında en düşük RMSE değerinin 0.30 SH durdurma kuralında elde edildiği görülmektedir. Yanlılık değerleri ve testte yer alan madde sayılarına ilişkin önemli farklılık görülmemektedir. Sonuç olarak, bu çalışmada, teste başlama kuralının, koşullara göre ölçme kesinliği ve test uzunluğu açısından farklılık oluşturmadığı görülmüştür. RMSE ve yanlılık değerlerinin daha düşük elde edildiği durdurma kuralının 0,30 SH durdurma kuralı olduğu sonucuna ulaşılmıştır.

Yetenek Kestirim Yöntemi Olarak MOK, MSD veya BSD Belirlenmesine İlişkin Sonuçlar

Madde seçim yöntemi MFB ve teste başlama kuralı θ=0 koşullarında tüm yetenek kestirim yöntemlerinde 0.30 SH durdurma kuralının kullanıldığı durumlarda, en düşük RMSE değerinin elde edildiği görülmektedir. Bununla birlikte, BSD yetenek kestirim yönteminden elde edilen RMSE ve yanlılık değerlerinin daha düşük olduğu görülmektedir. Yanlılık değerleri göz önünde bulundurulduğunda ise tüm koşullarda, MSD yetenek kestirim yönteminin kullanıldığı koşulda daha düşük değerler elde edildiği

görülmektedir. Testte uygulanan madde sayısı dikkate alındığında ise, en düşük test uzunluğunun yetenek kestirim yöntemi MSD seçildiğinde ve 0.50 SH durdurma kuralı kullanıldığında, en yüksek test uzunluğunun ise yetenek kestirim yöntemi MOK seçildiğinde 0.30 SH durdurma kuralında elde edildiği görülmektedir. Test uzunlukları dikkate alındığında ise durdurma koşullarında önemli farklılıklar elde edilmediği görülmektedir.

Madde seçim yöntemi KLB ve teste başlama kuralı θ=0 seçildiğinde, tüm yetenek kestirim yöntemlerinden elde edilen RMSE değerleri incelendiğinde, en düşük RMSE değerinin 0.30 SH durdurma kuralında elde edildiği görülmektedir. MOK yetenek kestirim yöntemi ile elde edilen RMSE ve yanlılık değerlerinin, diğer yetenek kestirim yöntemlerine göre daha düşük elde edildiği görülmektedir. MSD yetenek kestirim yönteminin kullanıldığı tüm koşullarda, daha düşük yanlılık değerlerinin elde edildiği gözlenmektedir. Test uzunlukları incelendiğinde, en düşük test uzunluğu değerinin, yetenek kestirim yöntemi MSD ve 0.50 SH durdurma kuralının olduğu durumda, en yüksek test uzunluğu değerinin ise yetenek kestirim yöntemi MOK ve 0.30 SH durdurma kuralı kullanıldığında gözlendiği görülmektedir. Test uzunlukları karşılaştırıldığında, durdurma kurallarının koşulları arasında önemli farklılıklar görülmediği sonucuna ulaşılmıştır.

Madde seçim yöntemi MFB ve teste başlama kuralı -1<θ<1 koşullarında; tüm yetenek kestirimleri için en düşük RMSE değerinin 0.30 SH durdurma kuralında elde edildiği görülmektedir. BSD yetenek kestirim yöntemi ile elde edilen RMSE ve yanlılık değerlerinin daha düşük elde edildiği görülmektedir. Yanlılık değerlerinin ise tüm koşullarda, MSD yetenek kestirim yönteminin kullanıldığı durumda daha düşük olduğu görülmektedir. Test uzunlukları incelendiğinde, en düşük test uzunluğu değerinin yetenek kestirim yöntemi MSD ve 0.50 SH durdurma kuralında, en yüksek test uzunluğu değerinin ise yetenek kestirim yöntemi MOK ve 0.30 SH durdurma kuralı uygulandığı koşullarda elde edildiği görülmektedir. Test uzunlukları incelendiğinde, durdurma koşullarına göre önemli farklılıklar elde edilmediği görülmektedir.

Madde seçim yöntemi KLB ve teste başlama kuralı -1<θ<1seçildiğinde; tüm yetenek kestirimleri için en düşük RMSE değerinin 0.30 SH durdurma kuralında elde edildiği görülmektedir. BSD kestirim yöntemi ile elde edilen RMSE ve yanlılık değerlerinin diğer yetenek kestirim yöntemlerine göre daha düşük olduğu gözlenmektedir. Yanlılık değerleri karşılaştırıldığında ise tüm koşullarda, MSD yetenek kestirim yönteminin kullanıldığı durumda daha düşük değerler aldığı görülmektedir. Test uzunlukları karşılaştırıldığında, en düşük test uzunluğu değerinin yetenek kestirim yöntemi MSD ve 0.50 SH durdurma kuralında en yüksek test uzunluğu değerinin ise yetenek kestirim yöntemi MOK seçildiğinde 0.30 SH durdurma kuralı koşullarında elde edildiği görülmektedir. Test uzunlukları incelendiğinde, durdurma koşulları arasında çok önemli farklılıklar olmadığı görülmektedir.

Genel sonuç olarak, BSD yetenek kestirim yöntemi kullanıldığı koşullarda MOK yetenek kestirim yöntemine kıyasla daha düşük RMSE ve yanlılık değerleri elde edilmiştir. Literatür incelendiğinde, MOK ve BSD yetenek kestirim yöntemlerinin RMSE ve yanlılık değerlerinin karşılaştırılması ile ilgili, bu araştırmanın sonucunu destekleyecek çalışmaların (Wang & Vispoel,1998; İşeri,2002) mevcut olduğu görülmektedir.

Madde Seçim Yöntemi Olarak MFB ve KLB Belirlenmesine İlişkin Sonuçlar

Yetenek kestirimi MOK ve teste başlama kuralı θ=0 seçildiğinde; durdurma kurallarının kullanıldığı koşullarda, en düşük RMSE değerinin 0.30 SH durdurma kuralında ve madde seçim yönteminin KLB olarak ele alındığı koşulda elde edildiği görülmektedir. Yanlılık değerleri incelendiğinde, en düşük yanlılık değerlerinin madde seçim yöntemi KLB olarak ele alındığında elde edildiği gözlendiği görülmektedir. Test uzunlukları karşılaştırıldığında, en yüksek test uzunluğu değerinin ise madde seçme yöntemi MFB ve 0.30 SH durdurma kuralı koşulunda en düşük test uzunluğu değerinin, madde seçim yöntemi KLB ve 0.50 SH durdurma kuralı koşulunda gözlendiği görülmektedir.

Yetenek kestirimi MSD ve teste başlama kuralı θ=0 seçildiğinde; en düşük RMSE değerinin 0.30 SH (yetenek) durdurma kuralı ve madde seçim yönteminin KLB olarak ele alındığı koşulda elde edildiği görülmektedir. Yanlılık değerleri incelendiğinde, en düşük yanlılık değerlerinin madde seçim yöntemi KLB olarak ele alındığında elde edildiği görülmektedir. Test uzunlukları karşılaştırıldığında, en düşük test uzunluğu değerinin madde seçim yöntemi MFB olduğunda, 0.50 SH durdurma kuralında elde edildiği

görülmektedir. Test uzunlukları incelendiğinde, durdurma koşullarında ve madde seçim yöntemleri arasında çok önemli farklılıklar gözlemlenmemiştir.

Yetenek kestirimi BSD ve teste başlama kuralı θ=0 seçildiğinde en düşük RMSE değerinin 0.30 SH durdurma kuralı ve madde seçim yöntemi KLB olarak ele alındığı koşulda elde edildiği görülmektedir. Yanlılık değerleri incelendiğinde, en düşük yanlılık değerlerinin madde seçim yöntemi MFB olarak ele alındığında elde edildiği gözlendiği görülmektedir. Test uzunlukları incelendiğinde, durdurma koşullarında ve madde seçim yöntemleri arasında çok önemli farklılıklar gözlemlenmemiştir.

Yetenek kestirimi yöntemi MOK, MSD ve BSD olarak ele alındığında ve teste başlama kuralı -1<θ<1 seçildiğinde ise en düşük RMSE değerinin 0.30 SH durdurma kuralı ve madde seçim yöntemi KLB olarak ele alındığı koşulda elde edildiği görülmektedir. Yanlılık değerleri incelendiğinde, en düşük yanlılık değerlerinin madde seçim yöntemi KLB olarak ele alındığında gözlendiği görülmektedir. Test uzunlukları incelendiğinde, durdurma koşullarında ve madde seçim yöntemleri arasında çok önemli farklılıklar gözlemlenmemiştir.

Çalışmada, KLB madde seçim yönteminin MFB'ye kıyasla daha düşük RMSE ve yanlılık değerlerine sahip olduğu görülmektedir. KLB ve MFB madde seçim yöntemlerinin RMSE ve yanlılık değerlerinin karşılaştırılması ile ilgili gerçekleştirilen çalışmaların (Linda,1996; Eggen,1999) sonucu ile benzer sonuçlara ulaşılmıştır.

Araştırmadan elde edilen sonuçlar göz önünde bulundurulduğunda aşağıdaki öneriler sunulabilir.

1. BBT uygulamalarında, başlama kuralı olarak θ=0, yetenek kestirim yöntemi olarak MOK seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık değerleri hem de testin kullanışlılığı bakımından yetenek kestiriminde SH<0.30 sonlandırma koşulu önerilebilir.
2. BBT uygulamalarında, başlama kuralı olarak θ=0, yetenek kestirim yöntemi olarak MSD seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık değerleri hem de testin kullanışlılığı bakımından yetenek kestiriminde benzer şekilde SH<0.30 sonlandırma koşulu önerilebilir.
3. BBT uygulamalarında, başlama kuralı olarak θ=0, yetenek kestirim yöntemi olarak BSD seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık değerlerinde hem de testin kullanışlılığı bakımından yetenek kestiriminde SH<0.30 SH<0.30 sonlandırma koşulu önerilebilir.
4. BBT uygulamalarında, başlama kuralı olarak -1<θ<1, yetenek kestirim yöntemi olarak MOK seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık değerleri hem de testin kullanışlılığı bakımından yetenek kestiriminde SH<0.30 sonlandırma koşulu önerilebilir.
5. BBT uygulamalarında, başlama kuralı olarak -1<θ<1, yetenek kestirim yöntemi olarak MSD seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık değerlerinde hem de testin kullanışlılığı bakımından yetenek kestiriminde benzer şekilde SH<0.30 SH<0.30 sonlandırma koşulu önerilebilir.
6. BBT uygulamalarında, başlama kuralı olarak -1<θ<1, yetenek kestirim yöntemi olarak BSD seçildiğinde ve her iki madde seçim yöntemi kullanıldığında hem RMSE ve yanlılık hem de testin kullanışlılığı bakımından yetenek kestiriminde SH<0.30 sonlandırma koşulu önerilebilir.
7. BBT uygulamalarında, madde seçim yöntemi olarak KLB yönteminin, yetenek kestirim yöntemi olarak BSD'nin olduğu koşullarda hem RMSE ve yanlılık değerleri hem de testin kullanışlılığı bakımından yetenek kestiriminde SH<0.30 sonlandırma koşulu önerilebilir.

Benzer bir çalışma farklı madde havuzu büyüklükleriyle gerçekleştirilebilir. Ayrıca, madde havuzunun özellikleri değiştirilerek durdurma kurallarının karşılaştırılması yapılabilir. Çalışmada maddelerin kullanım sıklıkları göz önünde bulundurulmamıştır. Maddelerin kullanım sıklıklarını da dikkate alan benzer çalışmalar gerçekleştirilebilir.

KAYNAKÇA

Babcock, B. & Weiss, D. J. (2009). *Termination criteria in computerized adaptive tests: variable-length cats are not biased.* Paper presented at The 2009 Conference on Computerized Adaptive Testing, Minnesota, USA.https://www.researchgate.net/publication/262674764_Termination_Criteria_in_Computerized_Adaptive_Tests_Do_Variable-Length_CATs_Provide_Efficient_and_Effective_Measurement

Babcock, B. ve Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: do variable-length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1–18. https://doi.org/10.7333/1212-0101001

Baker, F.B. & Kim, S.H. (2004). *Item response theory: Parameter estimation techniques.* Marcel Bekker Inc.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examaninee's ability. In

Lord, F.M. & Novick, M.R. (Eds.) *Statistical theories of mental test scores* (pp. 397-479) . Addison-Wesley.

Blais, J.& Raiche, G. (2002). *Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Paper presented at The International Objective Measurement Workshop International Objective Measurement Workshop, New Orleans, USA. https://pubmed.ncbi.nlm.nih.gov/21164229/

Blais, J. & Raiche, G. (2010). Features of the sampling distribution of the ability estimate in Computerized Adaptive Testing according to two stopping rules, *Journal of Applied Measurement, 11*(4), 424-31. https://www.researchgate.net/publication/49689146

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

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

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research,12*(49), 61-80. https://files.eric.ed.gov/fulltext/EJ1059924.pdf

Chang, S. W. & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*(1), 71–103. https://doi.org/10.1111/j.1745-3984.2003.tb01097.x

Chang, H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20* (3), 213–229. https://doi.org/10.1177/014662169602000303

Chang, H. & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 333-341. https://www.researchgate.net/publication/238681527

Choi, S. W., Grady, M.W., & Dodd, B.G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 70*(6), 1-17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3028267/

Deng, H., Ansley, T., & Chang, H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, 47(2), 202-226.

https://onlinelibrary.wiley.com/journal/17453984

Eggen, T. H. J. M. (1999). Item Selection in Adaptive Testing with the Squential Probability Ratio Test. *Applied Psychological Measurement, 23*(3), 249-261. https://doi.org/10.1177/01466219922031365

Eggen, T. (2004). *Contributions to the theory and practice of Computerized Adaptive Testing*. (Unpublished doctoral dissertation). University of Twente, Enschede, Netherlands.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Eroğlu, M. G. & Kelecioğlu, H. (2015). Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 28*(1), 31-52. https://doi.org/10.19171/uuefd.87973

Evans, J. J. (2010). *Comparability of examinee proficiency scores on computer adaptive tests using real and simulated data.* (Unpublished doctoral dissertation). The State University of New Jersey, New Brunswick, United States.

Glas, C.A. & Linden, W. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27* (4), 247–261. https://doi.org/10.1177/0146621603027004001

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and application.* Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of itemresponse theory*. Sage Publications Inc.

Hambleton, R. K. & Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19(3)*, 221–239. https://www.tandfonline.com/journals/hame20

Han, K. T. (2009). *A gradual maximum information ratio approach to item selectionin computerized adaptive testing*. Paper presented at The Conference on Computerized Adaptive Testing, Minnesota, USA. http://www.iacat.org/sites/default/files/biblio/cat09han.pdf

Han, K. T. (2010). *Comparision of non-fisher information item selection criteria in fixed length computerized adaptive testing.* Paper presented at The Annual Meeting of the National Council on Measurement in Education, Denver, USA.http://www.umass.edu/remp/software/simcata/papers/NCME2010_1_HAN.pdf

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *AppliedPsychological Measurement, 20*(2), 101–125. https://journals.sagepub.com/doi/10.1177/014662169602000201

Ho, T. (2010). A comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on generalized partial credit model. (Unpublished doctoral dissertation). The State University of Texas, TX, United States.

İşeri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testingprocedures.* (Yayımlanmamış doktora tezi). Orta Doğu Teknik Üniversitesi, Ankara, Türkiye.

Ivei, J. L. (2007). *Test taking strategies in computer adaptive testing that will ımprove your score: factor fiction?.* (Unpublished doctoral dissertation). The State University of Texas, TX, United States.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability*. (Yayımlanmamış doktora tezi). Orta Doğu Teknik Üniversitesi, Ankara, Türkiye.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359- 375. https://www.tandfonline.com/doi/abs/10.1207/s15324818ame0204_6?journalCode=hame20

Linda, T. (1996). *A comparision of the traditional maximum information method and the global information method in CAT item selection*. Paper presented at The Annual Meeting of the National Council on Measurement in Education, New York, NY USA. http://www.iacat.org/content/comparison-traditional-maximum-information-method-and-global-information-method-cat-item

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

Lord, F. M. (1983). Unbiased estimators of ability parameters of their variance, and of their parallel- forms reliability. *Psychometrika, 48*(2), 233-245. https://doi.org/10.1007/BF02294018

Lord, F. & Stocking, M. (1988). Item reponse theory. In J. P. Keeves (Eds.). *Educational research, methodology, and measurement: An international handbook* (pp. 269-272) . Pergamon Press.

MacDonald, P. L. (2002). *Computer adaptive test for measuring personality factors using item response theory.* (Unpublished doctoral dissertation). The University Western of Ontario, Ontario, Canada.

Magis, D. & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testingwith the R package catR. *Journal of Statistical Software, 48* (8), 1-31. https://www.jstatsoft.org/article/view/v048i08

McLeod, L. D. & Schnipke, D. L. (1999). *Detecting items that have been memorized in thecomputerized adaptive testing environment.* Paper presented at The Annual Meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada. https://files.eric.ed.gov/fulltext/ED432592.pdf

Mills, C. N. & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *AppliedMesurement in Education, 9*(4), 287-304. https://www.tandfonline.com/doi/abs/10.1207/s15324818ame0904_1?journalCode=hame20

Orcutt, V. L. (2002). *Computerized adaptive testing: Some issues in development.* Paper presented at TheAnnual Meeting of the Educational Research Exchange, Denton, TX USA. https://www.academia.edu/48173923/Computerized_Adaptive_Testing_Some_Issues_in_Development

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing*.New York, NY: Springer

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Eds.). *Newhorizonsin testing: latent trait theory and computerized adaptive testing*. New York: Academic Press

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihoodestimate of ability. *Psychometrika, 42*(2), *163-191.* https://doi.org/10.1007/BF02294047

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika,61*(2), 331-354. https://media.metrik.de/uploads/incoming/pub/Literatur/1996_Multidimensional%20adaptive%20testing.pdf

Segall, D. O. (2004). Computerized adaptive testing. In Kempf-Leanard (Eds.). *The encyclopedia ofsocial measurement,* (pp. 429 – 438). Academic Press.

Sereci, S. (2003). Computerized Adaptive Testing: An Introduction. In J.E. Wall ve G.R. Walz (Eds.).*Measuring Up: Assessment Issues for Teachers, Counselors and Administrators*, (pp.685-697). CAPS Press.

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

Simms, L. J. & Clark, L. A. (2005) . Validation of a computerized adaptive version of the schedulefor non-adaptive and adaptive personality (SNAP). *Psychological Assessment, 17*(1), 28-43. https://doi.org/10.1037/1040-3590.17.1.28

Spray, J. A. & Reckase, M. D. (1994). *The selection of test items for decision making with a computer adaptive test*. Paper presented at The Annual Meeting of the National Council on Measurement in Education. New Orleans, LA, United States. https://files.eric.ed.gov/fulltext/ED372078.pdf

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research

Report No. 93-2). Princeton, NJ: Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1993.tb01513.x

Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması*. (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi, Ankara, Türkiye.

Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. Journal of Measurement and Evaluation in Education and Psychology, 10(3), 315-326. https://dergipark.org.tr/tr/pub/epod

Şahin, A. & Özbaşı, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive testing. *Eurasian Journal of Educational Research,* 17(69), 21-36. http://dergipark.org.tr/ejer/issue/42462/511414

Tatsuoka, C. & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics, 65*, 143–157. https://www.jstor.org/stable/3088831

Thompson, N. A. (2007b). Computerized classification testing with composite hypotheses. Paper presented at The GMAC Conference on Computerized Adaptive Testing, Minneapolis, United States. https://www.researchgate.net/publication/229046974_Computerized_classification_testing_with_composite_hypotheses

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*(1), 1-9. http://pareonline.net/getvn.asp?v=16&n=1

Thissen, D. & Steinberg, L. (2009). Item response theory. In R. Millsap ve A. Maydeu-Olivares (Eds.) *The sage handbook of quantitative methods in psychology*. Sage Publications.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196. https://onlinelibrary.wiley.com/journal/17453984

Veldkamp, B.P. (2012). Ensurind The Future of Computerized Adaptive Testing. In Theo J.H.M. Eggen ve Veldkamp, B.P. (Eds.). *Psychometrics in Practice at RCEC*, (pp.39-50). RCEC.

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*(4), 575–588. https://doi.org/10.1007/BF02295132

Veldkamp, B.P. & van der Linden. W.J. (2010). Designing item pools for adaptive testing. In W.J van der Linden. ve C.A.W. Glas.(Eds.). *Computerized adaptive testing: Theory and practice*, (pp.149-162). Springer.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods computerized adaptive testing. *Journal of Educational Measurement, 35* (2), 109-135. https://www.jstor.org/stable/1435235

Wainer, H. (2000). *Computerized Adaptive Testing*. Lawrence Erlbaum Assc.

Wang, S., & Wang, T. (2001). Precision of warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement, 25*(4), 317-331. https://journals.sagepub.com/doi/10.1177/01466210122032163

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450. https://doi.org/10.1007/BF02294627

Wen, H., Chang, H., & Hau, K. (2000). *Adaption of a-stratified method in variable length computerized adaptive testing*. Paper presented at The American Educational Research Association Annual Meeting, Seattle, USA. https://eric.ed.gov/?id=ED465763

Weiss, D. J.(1982). Improving measurement quality and efficiency with Adaptive Testing. *Applied Psychological Mesurement, 6*(4),473-492. https://doi.org/10.1177/014662168200600408

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

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation inCounseling and Development, 37* (2), 70-84. https://doi.org/10.1080/07481756.2004.11909751

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

Weissman, A. (2003). *Assessing the efficiency of item selection in computerized adaptive testing.* (Unpublished doctoral dissertation), University of Pittsburgh, Pensilvanya, United States.

Yi, Q., Wang, T., & Ban, J.C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38*(3), 267-292. https://www.jstor.org/stable/1435124

Yi, Q., & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, 56 (2),359–378. https://pubmed.ncbi.nlm.nih.gov/14633340/