ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(1); 15-27

Investigation of Classification Accuracy, Test Length and Measurement Precision at Computerized Adaptive Classification Tests *

Seda DEMİR ** Burcu ATAR ***

Abstract
This study aims to compare Sequential Probability Ratio Test (SPRT) and Confidence Interval (CI) classification criteria, Maximum Fisher Information method on the basis of estimated-ability (MFI-EB) and Cut-Point (MFI-CB) item selection methods while ability estimation method is Weighted Likelihood Estimation (WLE) in Computerized Adaptive Classification Testing (CACT), according to the Average Classification Accuracy (ACA), Average Test Length (ATL), and measurement precision under content balancing (Constrained Computerized Adaptive Testing: CCAT and Modified Multinomial Model: MMM) and item exposure control (Sympson-Hetter Method: SH and Item Eligibility Method: IE) when the classification is done based on two, three, or four categories for a unidimensional pool of dichotomous items. Forty-eight conditions are created in Monte Carlo (MC) simulation for the data, generated in R software, including 500 items and 5000 examinees, and the results are calculated over 30 replications. As a result of the study, it was observed that CI performs better in terms of ATL, and SPRT performs better in ACA and correlation, bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) values, sequentially; MFI-EB is more useful than MFI-CB. It was also seen that MMM is more successful in content balancing, whereas CCAT is better in terms of test efficiency (ATL and ACA), and IE is superior in terms of item exposure control though SH is more beneficial in test efficiency. Besides, increasing the number of classification categories increases ATL but decreases ACA, and it gives better results in terms of the correlation, bias, RMSE, and MAE values.

Key Words: Computerized adaptive classification testing, content balancing, item exposure control, classification criteria, item selection methods.

INTRODUCTION

Testing in education might have various objectives. These objectives include increasing the effectiveness of education, assessing students individually, making selection or placement decisions, certification, monitoring learning progress, and testing for diagnostic purposes. To achieve these objectives, it seems to be critical to have access to timely and accurate information about learners' level of ability. In this regard, Computerized Adaptive Testing (CAT) is one of the greatest reflections of developments in information and communication technologies in the field of education and contributes to making more qualified and effective evaluations.

Unlike traditional paper-pencil tests, a CAT system uses different test forms in real time based on their individualized performance to test individuals with different levels of ability (Bao, Shen, Wang, & Bradshaw, 2021). The goal of CAT is to estimate each individual's latent ability and select the most appropriate test items (i.e., the most informative item) from the item pool for an individual based on his or her current performance (Eggen & Straetmans, 2000). At the end of the process, CAT provides more reliable estimates of ability using fewer items compared to traditional tests (Bao et al., 2021;

* This study is based on doctoral dissertation entitled "Investigation of Classification Accuracy at Computerized Adaptive Classification Tests"

** Assist. Prof., Tokat Gaziosmanpasa University, Faculty of Education, Tokat-Turkey, seddadmr@gmail.com, ORCID ID: 0000-0003-4230-5593
*** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Turkey, burcua@hacettepe.edu.tr, ORCID ID: 0000-0003-3527-686X

To cite this article:
Demir, S., & Atar, B. (2021). Investigation of classification accuracy, test length and measurement precision at computerized adaptive classification tests. *Journal of Measurement and Evaluation in Education and Psychology, 12*(1), 15-27. doi: 10.21031/epod.787865

Received: 30.08.2020
Accepted: 21.02.2021

Fan, Wang, Chang, & Douglas, 2012; Thompson, 2009). These advantages of CAT can be seen as the main reason for preferring large scale CAT applications such as the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE), and the National Assessment of Educational Progress (NAEP). The main purpose of testing individuals may sometimes be the accuracy of classifications, such as passed or failed, apart from the effective estimate of ability. In that case, a Computerized Adaptive Classification Test (CACT) is preferred. Since important decisions are made based on the classification (e.g., retention, high school graduation, career selection), efficient and accurate classification is of critical importance (Thompson & Ro, 2007).

Additionally, test effectiveness is important for both CATs and CACTs. High test effectiveness in CAT applications with a unidimensional item pool means fewer items and lower standard errors for ability estimation (van der Linden & Hambleton, 1996 as cited in Thompson, 2009). Unlike CATs, CACTs use as few items as possible and aim at low classification errors to achieve test effectiveness (Thompson, 2009).

Purpose of the Study

An extensive review of literature on CACT applications revealed that most of the studies considered classification in only two categories (e.g., Gündeğer & Doğan, 2018a; Lau, 1996; Reckase, 1983; Spray & Reckase, 1996), and content balancing and item exposure control were not taken into account. Furthermore, classification criteria (e.g., Kingsbury & Weiss, 1980; Spray & Reckase, 1996; Thompson, 2009) and item selection methods were mostly compared (e.g., Gündeğer & Doğan, 2018b; Eggen, 1999; Lin & Spray, 2000), and the performance of different item selection methods was examined by crossing the item selection methods with classification criteria (e.g., Eggen & Straetmans, 2000; Thompson & Ro, 2007). Besides, there are a few studies that compared the performance of classification criteria in terms of Average Classification Accuracy (ACA) and Average Test Length (ATL) according to different item exposure control methods (Huebner, 2012; Lau & Wang, 1999). A study used the Sympson-Hetter (SH) item exposure control method together with the spiral method for content balancing (Huebner & Li, 2012). Considering the contribution of accurate classifications to selecting, monitoring, or placing individuals based on the test results, there seems to be a need for new research in CACT using different research designs. It is thus thought that this study will contribute to a deeper understanding of CACT applications.

The main purpose of this study was to examine the performance of different classification criteria and item selection methods used in CACT applications when weighted likelihood estimation (WLE) is used for ability estimation under various conditions of classification category numbers, content balancing, and item exposure control methods in terms of average classification accuracy, average test length, the correlation between true and estimated ability levels, bias, root mean squared error (RMSE), and mean absolute error (MAE). The research problems are as follows:

Given that WLE is the ability estimation method, and the sequential probability ratio test (SPRT) with indifference region (IR) constant value δ: .20, and the confidence interval with CI: 90% confidence level are the classification criteria, how do the values of average classification accuracy, average test length, the correlation between true and estimated ability levels, bias, RMSE, and MAE change in two, three or four-category classifications where the followings are considered together?

1. The estimate-based maximum Fisher information (MFI-EB) and cut score-based maximum Fisher information (MFI-CB) item selection methods,
2. The MFI-EB and MFI-CB item selection methods along with the constrained CAT (CCAT) and modified multinomial model (MMM) content balancing methods, and the Sympson-Hetter (SH) and item eligibility (IE) item exposure control methods.

For the purpose of the research, below are described the design of the simulation study, data generation, CACT simulation conditions, and analysis plan. Then, the results are summarized, and the main findings are highlighted. Finally, a discussion is given on the implications of this simulation study according to ACA, ATL, measurement precision, and its results, and suggestions for future research.

METHOD

In this study, Monte Carlo (MC) simulations were performed, and CACT application results were compared using simulated datasets. If other research methods answer the questions What happened, and how, and why? simulation studies help answer the question What if ...? In simulation studies, it is possible to examine more complex systems as possible different conditions into the future can be created (Dooley, 2002). The datasets used were generated in the R program (R Core Team, 2013) based on the conditions examined in the study. The dependent variables of the study were ACA, ATL, correlation between real ability values and estimated ability values (*r*), bias, RMSE, and MAE. The independent variables were classification criteria (SPRT and CI), item selection methods (MFI-EB and MFI-CB), content balancing methods (CCAT and MMM), item exposure control methods (SH and IE), and the number of classification categories (two, three, and four). Therefore, the study had 48 simulation conditions = 2 classification criteria x 2 item selection methods x 2 content balancing methods x 2 item exposure control methods x 3 classification category numbers.

Data Generation

The data used in this study were generated by simulation in accordance with certain properties.

Generation of item and ability parameters for Monte Carlo (MC) simulation

This study was conducted as an MC simulation study by taking Thompson's (2011) study into consideration. The item pool was composed of 500 items under Item Response Theory (IRT) three-parameter logistic model (3PLM) for each of 30 replications. Since both estimate-based and cut score-based item selection methods (MFI-EB and MFI-CB) were used and two-, three- or four-category classifications were made, the item pool was composed of items that provide a high amount of information at and around the cut-point $\theta = 0$ and cover the ability level range (-3, 3). For the items in the pool, the a parameter was generated from a uniform distribution $U[0.5, 2.0]$ to represent medium and high levels of discrimination considering the study of Kingsbury and Weiss (1980), the *b* parameter was generated from a normal distribution $N(-0.5, 1.5)$ to be close to the actual values in applications as pointed out in Thompson (2009) and Warm (1989), and the *c* parameter was generated from a normal distribution $N(0.20, 0.05)$ again to be close to an actual application in keeping with Thompson (2009). In addition, ability parameters of 5000 examinees were generated from a normal distribution $N(0, 1)$ within a range of (-3, +3) for each of 30 replications.

CACT Simulation Conditions

CACT simulation conditions, used in this study, were explained in detail under subheadings.

Starting point

Available prior information about examinees can be used as the starting point in CACT (Weiss & Kingsbury, 1984; Yang, Poggio, & Glasnapp, 2006). Although not used very often, the population mean can also be defined as the starting point (Thompson, 2007b). In this research, the starting point for all conditions was determined as $\theta = 0$.

Item selection

Intelligent item selection methods where the computer program evaluates the unused items in the pool and decides which would be the best item to use next are generally classified into two groups: estimate-based and cut score-based (Thompson, 2007b). When IRT is used as the psychometric model, the cut score-based methods such as MFI, maximum Kullback-Leibler information (KLI), and log-odds ratio methods can be preferred (Lin & Spray, 2000). Traditionally, an item selection method that maximizes Fisher information at the cut-point is used with SPRT. SPRT is expected to yield better results, especially as the indifference region increases (Eggen, 1999). MFI-EB and MFI-CB methods were used for item selection in this study.

Ability estimation

Based on the literature, there are several ability estimation methods for binary scoring (1-0) and unidimensional item response theory modeling. The most common and widely used ability estimation methods include Maximum Likelihood Estimation (MLE), Marginal Maximum Likelihood Estimation (MMLE), Weighted Likelihood Estimation (WLE), and the Bayesian estimation methods such as Owen’s Bayesian sequential method, Maximum A Posteriori (MAP), and expected a posteriori (EAP). Warm (1989) noted that all these methods can produce some biased estimates. Bias affects the accuracy of classification decisions systematically (Wang & Wang, 2001). Additionally, Warm (1989) concluded that, especially in fixed-length tests, estimations made by WLE had less bias compared to estimations made by MLE and MAP. He discussed that when WLE is used for various lengths of adaptive tests, the test is similar to MAP but ends with fewer items than MLE, and he proposed the WLE method, which is a modified version of MLE, for ability estimation. This estimation method may reduce item exposure and test time, thereby enhancing the usefulness of the test. Thus, it can be considered as an advantage to use WLE for CACT and CAT applications. WLE is a method that reduces bias and works on the basis of item parameters and a weighting function specific to ability levels (Warm, 1989). WLE is most often preferred in CACT applications (Eggen & Straetmans, 2000; Nydick, Nozawa, & Zhu, 2012; Wouda & Eggen, 2009; Yang et al., 2006). Considering its advantages and its position in the literature about classification, WLE was used as an ability estimation method in this study. The WLE ability estimation method is a condition that was kept constant in simulations.

Classification criteria

There are three basic classification criteria based on IRT in CACT applications: SPRT, CI, and Bayesian decision theory. All three classification criteria require fewer items than traditional fixed-form tests and provide a similar level of classification accuracy (Kingsbury & Weiss, 1983). Previous research has shown that CI is more effective in estimate-based item selections, while SPRT is more effective in cutscore-based item selections (Eggen & Straetmans, 2000; Spray & Reckase, 1996; Thompson, 2009). It has also been shown that SPRT is more effective than CI, especially in terms of classification accuracy (Eggen, & Straetmans, 2000). Furthermore, as Thompson (2009) pointed out, the most used classification criterion in CACT studies is SPRT. Against this background, the classification criteria were determined as SPRT (δ: .20) and CI (90%) in this study.

Content balancing

In the content-balanced ICT applications, examinees are measured by a test that represents each of the content areas as appropriately as possible and has higher validity. The most commonly used content balancing methods in CACT studies are the spiralling method (Kingsbury & Zara, 1989) (e.g., Finkelman, 2008; Huebner, 2012) and the constrained CAT (CCAT) method (e.g., Eggen & Straetmans, 2000; Huebner & Li, 2012). Lin (2011) used a modified multinomial model (MMM) for content balancing. However, no research has been found that compares CCAT and MMM in the literature. Therefore, in this study, unlike the previous studies, two different content balancing methods, namely CCAT and MMM, were used. The minimum number of items to be used before terminating the test was set at 10, and the maximum number of items was set at 70 to ensure content balancing conditions. In cases where CCAT and MMM were included in the study conditions, the item pool generated with 500 items in the R program was divided into four content areas using random item assignment. Then, items were selected using the functions and loops written by the researcher in line with these content areas. The target proportions of four content areas were set at 40%, 30%, 20%, and 10%, respectively.

Item exposure

In CAT applications in which the item exposure control is not used, the selection of the items only based on maximum information could result in overexposure of items. On the other hand, both test security and more balanced use of item pool are considered while maintaining measurement precision when item exposure control techniques are implemented (Leroux et al., 2019). A search of the literature showed that the most used item exposure control methods in CACT applications are the random item selection method based on randomness strategies and the SH method (Sympson & Hetter, 1985) based on conditional selection strategies. Because randomness strategies are believed to be not effective under realistic test conditions, this research focused on the SH method and the IE method (van der Linden & Veldkamp, 2004), which is based on the same approach as the SH method. The maximum desired item exposure rate for the SH and IE methods used in the item exposure control was taken as $r_{\max} = .20$ (Leung, Chang, & Hau, 2002), which is a frequently used value in line with the studies of Huebner (2012) and Huebner and Li (2012).

Number of classification categories

Much of the research in CACT so far has used only two categories, such as failed-passed and a single cut-point. A two-category classification such as failed-passed was used in Huebner (2012), Lin and Spray (2000), Reckase (1983), Sie, Finkelman, Riley, and Smits (2015), Thompson (2009), van Groen, Eggen, and Veldkamp (2016). Both two- and three-category classifications were used in Eggen (1999) and Thompson (2007a). A three-category classification was used in Nydick et al. (2012). Both three- and five-category classifications were used in Yang et al. (2006). This research used two-, three- and four-category classifications to compare the changes. The ability parameters generated in R for the examinees were utilized to determine the cutting points for the classifications. The generated ability parameters were ranked from the low ability level to the high ability level. Through the method used in Eggen and Straetmans (2000), a cut-point was determined for the two-category classification, two cut-points were determined for the three-category classification, and three cut-points were determined for the four-category classification. In the two-category classification, the first half of the skill levels ranked from low to high were coded as Level 1 and the second half as Level 2. Then, the cut-point (CP = 0.00) was determined by taking 70% of the highest ability level in Level 1. Similarly, in the three-category classification, the ranked ability levels were encoded as Level 1, Level 2, and Level 3, and the cut-points were defined as CP1 = -0.29 and CP2 = 0.31. In the four-category classification, the ability levels were encoded as Level 1, Level 2, Level 3, and Level 4 and the cut-points were defined as CP1 = -0.47, CP2 = -0.01, and CP3 = 0.48.

Data Analysis

Thirty replications were conducted for each of the 48 simulation conditions generated within the scope of the research, and the values of the dependent variables were obtained by calculating the average of the replications. The value of the correlation between true and estimated ability levels was calculated using the Pearson correlation coefficient (PCC), while the bias, RMSE, and MAE values were calculated following formulas written in the R program.

Bias is calculated using the formula below where the sum of the difference between the last estimated ability level ($\widehat{\theta_\iota}$) and the true ability level (θ_i) is divided by the number of examinees (n) (Miller, & Miller, 2004):

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_\iota} - \theta_i)}{n}$$

RMSE is equal to the square root of the sum of squared of differences between the $\widehat{\theta_\iota}$ and θ_i divided by n:

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

MAE is calculated by dividing the sum of the absolute value of the difference between $\widehat{\theta_\iota}$ and θ_i by n:

$$OMH = \frac{\sum_{i=1}^{n}|\widehat{\theta_\iota} - \theta_i|}{n}$$

Additionally, functions and loops were written in the R program in addition to the item selection method for content balancing and item exposure control.

RESULTS

The results obtained for each subproblem of the study are presented under subheadings.

Results on the First Subproblem

Table 1 shows the values calculated by averaging 30 replications performed for each simulation condition related to the first research subproblem.

Table 1. Comparison of the Classification Criteria (CC) and Item Selection Methods (ISM) According to the Average Test Length (ATL), Average Classification Accuracy (ACA), and Measurement Precision With Correlation (r), Bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) Values When the number of Classification Categories (NCC) Based on Two, Three, or Four

CC	ISM	NCC	ATL	ACA	*r*	Bias	RMSE	MAE
SPRT (*δ* = .20)	MFB-EB	Two	24.72	.94	.94	-0.011	0.35	0.27
		Three	34.08	.88	.96	-0.012	0.32	0.24
		Four	41.34	.82	.96	-0.014	0.29	0.22
	MFB-CB	Two	22.95	.94	.90	0.019	0.44	0.32
		Three	33.93	.89	.92	0.015	0.38	0.28
		Four	42.88	.82	.93	0.012	0.35	0.26
CI (90%)	MFB-EB	Two	11.33	.89	.90	0.016	0.46	0.35
		Three	12.52	.79	.91	0.015	0.45	0.35
		Four	13.81	.71	.91	0.016	0.44	0.34
	MFB-CB	Two	11.55	.90	.87	0.019	0.49	0.38
		Three	12.62	.80	.87	0.017	0.48	0.37
		Four	13.82	.71	.88	0.020	0.47	0.36

Note. SPRT= sequential probability ratio test, CI= confidence interval, MFI-EB= maximum fisher information method on the basis of estimated-ability, MFI-CB= maximum fisher information method on the basis of cut-point.

As seen in Table 1, in the two-, three- and four-category classifications, the ACA values were quite high and ranged from .82 to .94, and the ATL values ranged from 22.95 to 42.88 when SPRT was used for classification. On the other hand, when CI was used for classification, the ACA values were relatively lower and ranged from .71 to .90, and the ATL values ranged from 11.33 to 13.82.

Accordingly, SPRT yielded better results in terms of ACA, and CI yielded better results in terms of ATL.

When the item selection methods MFI-EB and MFI-CB were used with the same classification criteria, similar results were obtained in terms of test effectiveness. In addition, an increase in the number of classification categories caused the test effectiveness to decrease for both classification criteria. In other words, it increased the ATL but reduced the ACA.

The values of the correlation (*r*) between the examinees' estimated and true ability levels ranged from .90 to .96 for SPRT and .87 to .91 for CI. With respect to the conditions in which the classification criteria were crossed by the item selection methods, higher correlations were calculated for both classification criteria in the conditions in which MFI-EB was used compared to the conditions in which MFI-CB was used. Additionally, similar correlation values were obtained in response to the increase in the number of classification categories. The bias calculated for the condition where SPRT and MFI-EB were used together (ranging from -0.014 to -0.011) was lower compared to that calculated for the condition where SPRT and MFI-CB were used together (ranging from 0.012 to 0.019). Similarly, the bias calculated for the condition where CI and MFI-EB were used together (ranging from 0.015 to 0.016) was lower compared to that calculated for the condition where CI and MFI-CB were used together (ranging from 0.017 to 0.020). The case is similar for the RMSE value, which takes into account the standard error of the estimation along with the bias, and for the MAE value. Accordingly, it can be said that lower bias, RMSE, and MAE values were found when the SPRT classification criterion or the MFI-EB item selection method was used. Furthermore, the increase in the number of categories did not exert a great effect on the bias but relatively decreased the RMSE and MAE values.

Results on the Second Subproblem

Table 2 demonstrates the values calculated by averaging 30 replications performed for each condition related to the second research subproblem, which incorporated CCAT and MMM for content balancing and SH and IE for item exposure control.

As seen in Table 2, in all conditions where the MMM content balancing method was used, the used content rates achieved the desired content rates (40%, 30%, 20%, and 10%, respectively). In the conditions where the CCAT content balancing method was used, the used content rates were above or below the desired content rates. For example, as seen in Table 2, in the condition where SPRT was used with MFI-CB, item exposure was controlled using IE, and a four-category classification was made, the CCAT content rates were found to be approximately 32%, 28%, 23%, and 16%, respectively. In addition, in the conditions where the IE item exposure control method was used, the proportion of items overexposed (OEX) was lower and the mean exposure rate of overexposed items (MOEX) achieved the desired $r_{\max} = .20$. On the other hand, in the conditions where SH was used, OEX was higher, and MOEX was considerably higher than the desired $r_{\max}= .20$. For example, as seen in Table 2, when SPRT and MFI-EB were used together, content balancing was done using CCAT, and a four-category classification was made, the OEX value calculated for item exposure controlled using SH was approximately .25, and the MOEX value was .29. In other words, approximately 25% of the items were above the maximum item exposure rate ($r_{\max} = .20$), and the mean item exposure was calculated to be approximately .29.

As seen in Table 2, another comparison using the same classification criteria and item selection method showed that although the CCAT content balancing method performed better with a slight difference in terms of test effectiveness, it generally produced similar results to MMM. In addition, the SH item exposure control method performed better compared to IE in terms of test effectiveness. The best result in terms of ATL (ATL = 11.13 and ACA = .88) was recorded in the condition where CI, MFI-EB, CCAT, and SH were used together, and a two-category classification was made, while the worst result (ATL = 51.93 and ACA = .75) was recorded in the condition where SPRT, MFI-CB, MMM, and IE were used together, and a four-category classification was made. To put it differently, it can be said that among the best and worst results, ATL was nearly five times higher, while ACA declined considerably.

Table 2. Comparison of The Classification Criteria (CC) and Item Selection Methods (ISM) According to the Average Test Length (ATL), Average Classification Accuracy (ACA), and Measurement Precision With Correlation (R), Bias, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) Values Under Content Balancing Methods (CBM) With Applied Content Rates and Item Exposure Control Methods (IECM) With Proportion of Items Overexposed (OEX), and Mean Exposure Rate of Overexposed Items (MOEX) When the Number of Classification Categories (NCC) Based on Two, Three, or Four

CC	ISM	CBM	IECM	NCC	Applied Content Rates				OEX	MOEX	ATL	ACA	*r*	Bias	RMSE	MAE
SPRT	MFB-EB	CCAT	SH	Two	36.27	29.61	21.78	12.34	.14	.28	26.91	.94	.94	-0.014	0.36	0.28
(δ= .20)			$r_{\max}$ = .20	Three	34.34	28.95	22.56	14.15	.21	.29	37.50	.87	.95	-0.015	0.32	0.24
				Four	33.32	28.65	22.94	15.09	.25	.29	44.69	.81	.96	-0.017	0.30	0.22
			IE	Two	35.56	29.38	22.09	12.97	.09	.20	29.87	.93	.94	-0.017	0.37	0.28
			$r_{\max}$ = .20	Three	33.71	28.77	22.79	14.74	.15	.20	41.71	.86	.95	-0.018	0.33	0.25
				Four	32.81	28.47	23.15	15.57	.18	.20	48.50	.78	.96	-0.018	0.31	0.23
		MMM	SH	Two	39.82	29.98	20.09	10.10	.14	.28	27.42	.94	.94	-0.015	0.37	0.28
			$r_{\max}$ = .20	Three	39.91	29.98	20.04	10.07	.21	.29	37.86	.87	.95	-0.015	0.33	0.25
				Four	39.92	29.99	20.05	10.04	.26	.29	45.35	.80	.96	-0.017	0.30	0.23
			IE	Two	39.80	30.03	20.06	10.11	.10	.20	30.82	.93	.94	-0.015	0.37	0.28
			$r_{\max}$ = .20	Three	39.84	30.01	20.08	10.08	.15	.20	42.27	.85	.95	-0.016	0.33	0.25
				Four	39.86	30.01	20.07	10.07	.17	.20	49.01	.77	.96	-0.018	0.32	0.24
	MFB-CB	CCAT	SH	Two	37.00	29.71	21.46	11.83	.13	.33	25.70	.94	.90	0.009	0.44	0.33
			$r_{\max}$ = .20	Three	34.41	28.98	22.49	14.12	.21	.34	38.35	.87	.92	0.009	0.39	0.28
				Four	32.96	28.56	23.07	15.41	.25	.36	47.02	.79	.93	0.006	0.36	0.26
			IE	Two	35.83	29.40	21.96	12.81	.11	.20	30.55	.93	.90	0.002	0.44	0.33
			$r_{\max}$ = .20	Three	33.46	28.74	22.90	14.91	.19	.20	43.94	.84	.92	0.003	0.40	0.30
				Four	32.40	28.39	23.30	15.92	.23	.20	51.14	.75	.93	0.000	0.38	0.28
		MMM	SH	Two	39.81	29.96	20.12	10.11	.13	.33	26.18	.94	.90	0.009	0.43	0.33
			$r_{\max}$ = .20	Three	39.88	30.00	20.08	10.04	.21	.34	38.61	.87	.92	0.008	0.39	0.29
				Four	39.92	30.01	20.03	10.05	.25	.36	47.36	.79	.93	0.006	0.36	0.26
			IE	Two	39.75	29.95	20.16	10.13	.12	.20	31.01	.93	.90	0.005	0.44	0.33
			$r_{\max}$ = .20	Three	39.83	30.00	20.07	10.10	.19	.20	44.85	.84	.92	0.003	0.40	0.30
				Four	39.84	30.01	20.05	10.09	.23	.20	51.93	.75	.93	0.006	0.39	0.29

(continued)

Table 2 (continue)

CC	ISM	CBM	IECM	NCC	Applied Content Rates				OEX	MOEX	ATL	ACA	r	Bias	RMSE	MAE
CI	MFB-EB	CCAT	SH	Two	42.13	30.95	19.09	7.83	.04	.27	11.13	.88	.89	0.011	0.48	0.37
90%			r_{max} = .20	Three	41.60	30.85	19.33	8.22	.05	.27	12.24	.78	.90	0.014	0.48	0.37
				Four	41.01	30.69	19.64	8.66	.05	.27	13.37	.70	.90	0.013	0.47	0.36
			IE	Two	42.04	30.90	19.17	7.90	.03	.20	11.19	.88	.89	0.016	0.49	0.38
			r_{max} = .20	Three	41.56	30.80	19.40	8.24	.03	.20	12.25	.78	.89	0.013	0.49	0.37
				Four	41.03	30.70	19.64	8.63	.03	.20	13.39	.69	.90	0.014	0.48	0.37
		MMM	SH	Two	39.97	30.04	19.96	10.04	.04	.27	11.18	.88	.89	0.015	0.48	0.37
			r_{max} = .20	Three	39.95	30.01	20.01	10.02	.05	.27	12.28	.78	.90	0.011	0.47	0.36
				Four	40.07	29.93	20.00	9.99	.05	.27	13.45	.70	.90	0.012	0.47	0.36
			IE	Two	40.06	29.94	20.01	9.99	.03	.20	11.21	.88	.89	0.012	0.49	0.38
			r_{max} = .20	Three	40.05	30.01	19.95	9.98	.03	.20	12.31	.78	.90	0.012	0.48	0.37
				Four	40.01	30.00	19.99	9.99	.03	.20	13.51	.69	.90	0.013	0.48	0.37
	MFB-CB	CCAT	SH	Two	41.96	30.93	19.19	7.92	.05	.36	11.36	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Three	41.49	30.77	19.40	8.34	.06	.36	12.46	.79	.87	0.01	0.49	0.39
				Four	40.88	30.64	19.68	8.80	.06	.36	13.69	.70	.87	0.01	0.48	0.38
			IE	Two	42.02	30.90	19.13	7.96	.05	.20	11.42	.88	.85	0.009	0.52	0.41
			r_{max} = .20	Three	41.43	30.75	19.46	8.37	.05	.20	12.56	.77	.86	0.004	0.51	0.40
				Four	40.55	30.55	19.86	9.03	.05	.20	14.6	.66	.86	0.006	0.51	0.39
		MMM	SH	Two	39.97	30.01	20.00	10.01	.06	.35	11.37	.89	.86	0.012	0.50	0.39
			r_{max} = .20	Three	39.91	30.02	20.07	9.99	.06	.35	12.56	.79	.87	0.009	0.49	0.38
				Four	39.97	30.02	20.00	10.01	.07	.35	13.65	.69	.87	0.013	0.49	0.38
			IE	Two	40.00	30.02	19.99	9.98	.05	.20	11.49	.88	.86	0.006	0.52	0.41
			r_{max} = .20	Three	40.04	29.98	20.01	9.97	.05	.20	12.74	.78	.86	0.008	0.51	0.40
				Four	39.99	29.98	20.02	10.00	.05	.20	14.42	.67	.87	0.009	0.50	0.39

Note: SPRT= sequential probability ratio test, CI= confidence interval, MFI-EB= maximum fisher information method on the basis of estimated-ability, MFI-CB= maximum fisher information method on the basis of cut-point, CCAT= constrained computerized adaptive testing, MMM= modified multinomial model, SH= Sympson-Hetter method, IE= item eligibility method and r_{max}= maximum desired item exposure rate.

The correlation (r) values ranged from .90 to .96 in the conditions where SPRT was used, while they ranged from .85 to .90 in the conditions where CI was used. The bias values ranged from -0.018 to

0.009 in the conditions where SPRT was used, while they ranged from 0.004 to 0.016 in the conditions where CI was used. The highest RMSE value (0.52) and the highest MAE value (0.41) were observed when CI, MFI-CB, CCAT (or MMM), and IE were used together, and a two-category classification was made. On the other hand, the lowest RMSE value (0.30) was observed when SPRT, MFI-EB, CCAT (or MMM), and SH were used together with four-category classification, and the lowest MAE value (0.22) was observed when SPRT, MFI-EB, CCAT, and SH were used together with four-category classification.

In summary, parallel to the findings in Table 1, CI performed better in terms of ATL, while SPRT performed better in terms of ACA. As the number of classification categories increased, ATL increased but ACA decreased. With respect to the correlation (*r*), bias, RMSE, and MAE values, SPRT performed better than CI, and MFI-EB performed better than MFI-CB. Furthermore, in response to the increased number of categories, the correlation and bias resulted in similar values, while the RMSE and MAE values were relatively lower.

DISCUSSION and CONCLUSION

Because the primary focus of this study is on classification accuracy, the ACA values calculated under different conditions are of great importance in interpreting the findings. In line with the research findings, high ACA values were calculated under all research conditions. The SPRT classification criterion performed better than CI and achieved a higher rate of classifying examinees into the accurate categories. On the other hand, the CI classification criterion performed better in terms of ATL under all research conditions and required fewer items to classify examinees compared to SPRT. This finding is in agreement with those obtained by Gündeğer and Doğan (2018a), Nydick et al. (2012), Thompson (2009), and Thompson and Ro (2007). These studies, in general, reported that the classifications made using CI ended with lower ATL and ACA compared to those made using SPRT. Therefore, comparing the SPRT and CI classification criteria used in the research in terms of classification accuracy, it may be suggested to prefer SPRT which yielded higher ACA values. On the other hand, comparing SPRT and CI in terms of ATL, CI seems to be preferable as it requires fewer items to classify examinees and terminate the test. Nevertheless, it should be noted that with respect to high-risk tests (e.g., tests applied in the field of medicine and directly related to human life), it is of key importance to choose the method which achieves a higher classification accuracy despite the increasing number of items. In CACTs, ATL, and ACA are often evaluated together for test effectiveness. If a decision is to be made to choose the best performing classification criterion in terms of test effectiveness, it may be suggested to use CI for conditions where both classification criteria achieve a good level of classification accuracy.

This research found that the SPRT classification criterion performed better than CI, and the MFI-EB item selection method performed better than MFI-CB in terms of measurement precision. Accordingly, under the conditions where the SPRT classification criterion or the MFI-EB item selection method was used, the values of correlation between examinees' true and estimated ability levels were higher while the bias, RMSE, and MAE values were lower. It can thus be said that examinees' last ability levels were more precise and closer to their true ability levels when the classification criterion was SPRT or when the item selection method was MFI-EB. A possible explanation of this result might be that the item pool was composed of items that provide great information at and around the cutting point $\theta = 0$. Additionally, the MFBI-EB item selection method achieved relatively better results compared to MFI-CB in terms of test effectiveness. In other words, when MFBI-EB was used, lower ATL values and similar ACA values were obtained.

The analysis results showed that the values of correlation between examinees' true and estimated ability levels were quite high, especially when the WLE ability estimation method was used together with the SPRT classification criterion and the MFI-EB item selection method. It can thus be said that the WLE method performs successfully.

Comparing the findings presented in Table 1 and Table 2, it can be seen that relatively higher ATL and lower ACA values were obtained in line with expectations when content balancing and item exposure control were added to the research conditions. According to Thompson (2007b), content balancing and item exposure constraints generally lead to an increase in only ATL. When content balancing and item exposure control are performed in CACT applications, it can be interpreted that the increase in ATL and the decrease in ACA may be due to the absence of an item that provides sufficient information about an examinee in the applied content area and does not exceed the item exposure rate. To solve this problem, the item pool might be expanded by increasing the number of items in each content area within the ability range which has plenty of items that exceed the maximum item exposure rate. The content balancing and item exposure control methods included in the research conditions did not change the correlation between examinees' true and estimated ability levels but caused a decrease in the bias values and an increase in RMSE and MAE values. The results obtained by the CI classification criterion were also little affected. This can be interpreted as an advantage provided by CI.

The research found that the MMM content balancing method performed better in achieving the desired content rates compared to CCAT. On the other hand, with respect to test effectiveness, CCAT performed better, especially in terms of ATL when SPRT was used although there were slight changes when CI was used. This finding is consistent with that reported by Lin (2011). Lin (2011) emphasized that although CCAT is one of the most chosen content balancing methods in CACTs, the MMM method, which is used mostly in CATs, is more successful in achieving the desired content balance. Therefore, in CACTs it is suggested to use MMM if content balancing is more critical as in high-risk tests, and CCAT if test effectiveness is more critical. The research also found that the IE method performed better in controlling item exposure compared to the SH method. This finding is in line with the work of Huebner (2012). Huebner (2012) concluded that IE works more successfully than SH in terms of item exposure control. In terms of test effectiveness, SH performed better, especially under the conditions where the SPRT classification criterion was used. When the SH method was used, lower ATL and higher ACA values were obtained. Thus, IE might be used if item exposure control, namely the safety of the test/item pool, is of critical importance in CACTs. Whereas SH might be used if test effectiveness is of more critical importance.

Under all research conditions, the increasing number of categories increased ATL while reducing ACA. To put it differently, the increasing number of categories reduced test effectiveness. This finding supports earlier observations in Eggen (1999) and Nydick et al. (2012). Eggen (1999) compared two-category and three-category classifications, and Nydick et al. (2012) compared three-category and five-category classifications. They found that the higher the number of categories was the higher the ATL values and the lower the ACA values were; thus, test effectiveness decreased. Therefore, in terms of test effectiveness, it may be suggested to keep the number of classification categories as few as possible. In addition, despite the increase in the number of classification categories, the correlation and bias values were similar, while RMSE and MAE values were relatively lower. Accordingly, examinees' last ability levels were more precisely estimated because the number of items required to terminate the test increased with the increasing number of classification categories. Therefore, it seems that the number of classification categories might be determined more optimally by considering correlation, bias, RMSE, and MAE values.

Based on the research findings, the following suggestions might be offered for future practice. If the focus of CACT is on ACA and content balancing and item exposure control are of critical importance, the SPRT classification criterion, which also performs better in terms of correlation, bias, RMSE, and MAE values, might be used together with the MFI-EB item selection method, the MMM content balancing method, and the IE item exposure control method. If the focus of CACT is on ATL and content balancing and item exposure control are performed, the CI classification criterion might be used together with MFI-EB, MMM, and IE. As for the researchers, in similar BBST studies, it can be recommended to use item pools with different properties such as multi-dimensional item pool or different pool sizes, skewness, kurtosis, etc. In addition, in similar studies to be conducted, the performances of the main BBST components can be compared over real data.

REFERENCES

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2021). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. doi: 10.1177/0146621620965730

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). London: Blackwell.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261. doi: 10.1177/01466219922031365

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713-734. doi: 10.1177/00131640021970862

Fan, Z., Wang, C., Chang, H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655-670. doi: 10.3102/1076998611422912

Finkelman, M. (2008). On using stochastic curtailment to shorten the SPRT in sequential mastery testing. *Journal of Educational and Behavioral Statistics, 33*(4), 442-463. doi: 10.3102/1076998607308573

Gündeğer, C., & Doğan, N. (2018a). A comparison of computerized adaptive classification test criteria in terms of test efficiency and measurement precision. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 161-177. doi: 10.21031/epod.401077

Gündeğer, C., & Doğan, N. (2018b). The effects of item pool characteristics on test length and classification accuracy in computerized adaptive classification testings. *Hacettepe University Journal of Education, 33*(4), 888-896. doi: 10.16986/HUJE.2016024284

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17*(12), 1-9. Retrieved from https://pareonline.net/getvn.asp?v=17&n=12

Huebner, A., & Li, Z. (2012). A stochastic method for balancing item exposure rates in computerized classification tests. *Applied Psychological Measurement, 36*(3), 181-188. doi: 10.1177/0146621612439932

Kingsbury, G. G., & Weiss, D. J. (1980). *A Comparison of adaptive, sequential and conventional testing strategies for mastery decisions* (Research Report 80-4). University of Minnesota, Minneapolis: MN. Retrieved from http://iacat.org/sites/default/files/biblio/ki80-04.pdf

Kingsbury, G. G., & Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing,* (pp. 237-254). New York: Academic Press.

Kingsbury, G. G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. doi: 10.1207/s15324818ame0204_6

Lau, C. A. (1996). *Robustness of a unidimensional computerized testing mastery procedure with multidimensional testing data* (Unpublished doctoral dissertation). University of Iowa, Iowa City IA.

Lau, C. A., & Wang, T. (1999, April). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Montreal, Canada. Retrieved from http://iacat.org/sites/default/files/biblio/la99-01.pdf

Leroux, A. J., Waid-Ebbs, J. K., Wen, P-S., Helmer, D. A., Graham, D. P., O'Connor, M. K, & Ray, K. (2019). An investigation of exposure control methods with variable-length cat using the partial credit model. *Applied Psychological Measurement, 43*(8),624-638. doi: 10.1177/0146621618824856

Leung, C.-K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson–Hetter algorithm. *Applied Psychological Measurement, 26*(4), 376-392. doi: 10.1177/014662102237795

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. *Applied Psychological Measurement 71*(1), 20-36. doi: 10.1177/0013164410387336

Lin, C. J., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. ACT (Research Report 2000-8). Iowa city, IA: ACT Research Report Series. Retrieved from https://eric.ed.gov/?id=ED445066

Miller, I., & Miller, M. (2004). *John E. Freund's mathematical statistics with applications*. (7th Ed.). New Jersey: Prentice Hall.

Nydick, S. W., Nozawa, Y., & Zhu, R. (2012, April). *Accuracy and efficiency in classifying examinees using computerized adaptive tests: An application to a large-scale test.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, British Columbia, Canada. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.3381&rep=rep1&type=pdf

R Core Team (2013). *R: A language and environment for statistical computing*, (Version 3.0.1) [Computer software], Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: latent trait theory and computerized adaptive testing*, (pp. 237-254). New York: Academic Press.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement, 39*(5), 389-405. doi: 10.1177/0146621615569504

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*(4), 405-414. doi: 10.3102/10769986021004405

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center. Retrieved from http://www.iacat.org/content/controlling-item-exposure-rates-computerized-adaptive-testing

Thompson, N. A. (2007a). *A comparison of two methods of polytomous computerized classification testing for multiple cutscores* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis.

Thompson, N. A. (2007b). A practitioner’s guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1), 1-13. Retrieved from http://www.iacat.org/sites/default/files/biblio/th07-01.pdf

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*(5), 778-793. doi: 10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4), 1-7. Retrieved from https://pareonline.net/getvn.asp?v=16&n=4

Thompson, N. A., & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC conference on computerized adaptive testing*. Retrieved from http://www.iacat.org/sites/default/files/biblio/cat07nthompson.pdf

Van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*(3), 273-291. doi: 10.3102/10769986029003273

Van Groen, M. M., Eggen, T. J. H. M., & Veldkamp, B. P. (2016). Multidimensional computerized adaptive testing for classifying examinees with within-dimensionality. *Applied Psychological Measurement, 40*(6), 387-404. doi: 10.1177/0146621616648931

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. doi: 10.1177/01466210122032163

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

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

Wouda, J. T., & Eggen, T. J. H. M. (2009). Computerized classification testing in more than two categories by using stochastic curtailment. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing*. Retrieved from http://iacat.org/sites/default/files/biblio/cat09wouda.pdf

Yang, X., Poggio, J. C., & Glasnapp, D. R. (2006). Effects of estimation bias on multiple category classification with an IRT-based adaptive classification procedure. *Educational and Psychological Measurement, 66*(4), 545-564. doi: 10.1177/0013164405284031