Journal of Educational Technology & Online Learning
Volume 5 | Issue 3 | 2022
http://dergipark.org.tr/jetol

The effect of item pool and selection algorithms on computerized classification testing (CCT) performance

Seda Demir [a] *

[a] Tokat Gaziosmanpasa University, Türkiye

Suggested citation: Demir, S. (2022). The effect of item pool and selection algorithms on computerized classification testing (CCT) performance. *Journal of Educational Technology & Online Learning*, *5*(3), 573-584.

Article Info

Keywords:

Computerized classification testing
Content balancing
Item exposure control
Classification accuracy
Item pool characteristics

Research Article

Abstract

The purpose of this research was to evaluate the effect of item pool and selection algorithms on computerized classification testing (CCT) performance in terms of some classification evaluation metrics. For this purpose, 1000 examinees' response patterns using the R package were generated and eight item pools with 150, 300, 450, and 600 items having different distributions were formed. A total of 100 iterations were performed for each research condition. The results indicated that average classification accuracy (ACA) was partially lower, but average test length (ATL) was higher in item pools having a broad distribution. It was determined that the observed differences were more apparent in the item pool with 150 items, and that item selection methods gave similar results in terms of ACA and ATL. The Sympson-Hetter method indicated advantages in terms of test efficiency, while the item eligibility method offered an improvement in terms of item exposure control. The modified multinomial model, on the other hand, was more effective in terms of content balancing.

1. Introduction

Computerized adaptive testing (CAT) has an increasing popularity among researchers and practitioners and has been widely preferred in the large-scale testing applications. Given that computers work more efficiently with item response theory (IRT), it can be noted that IRT is the fundamental theory in CAT applications. The most significant advantage of IRT is possibly the independence of item parameters across sub-samples and individual parameters from the items (Hambleton & Swaminathan, 1985). Although the examinees are tested through different sets of items, their scores can be compared with those of others thanks to these advantages of CAT applications as the items have been calibrated by IRT models (Krabbe, 2017). In this sense, CAT applications differ from the current paper and pencil tests in which all items are administered collectively in a fixed-form test. Through CAT, each examinee is offered a smaller set of items based on their individual performances in line with their ability levels. Thus, examinees' ability levels are more reliably estimated through CAT measurements (Bao, Shen, Wang, & Bradshaw, 2020; Krabbe, 2017; Fan, Wang, Chang, & Douglas, 2012; Thompson, 2009). On the other hand, computerized classification testing (CCT) is employed when the aim of the test is to classify participants based on test results in two or more categories. Considering the high stakes tests in such fields as medical licensure or educational proficiency examinations, the decision to be made according to the test type is directly related to people's life and future. Therefore, it is of great importance for CCT applications themselves that they have high classification accuracy rates (Thompson & Ro, 2007). Further, the fact that classification of

* Corresponding author. Department of Measurement and Evaluation in Education, Tokat Gaziosmanpasa University, Türkiye.
e-mail address: seddadmr@gmail.com

Doi: http://doi.org/10.31681/jetol.1099580
Received 6 Apr 2022; Revised 11 Jun 2022; Accepted 30 Jun 2022
ISSN: 2618-6586 .

examinees has been made through fewer items but high classification accuracy (lower classification errors) shows that CCT application has high level of test efficiency (Thompson, 2009). Weiss and Kingsbury (1984) note that there must be decisions to be made related to the components of a CAT procedure as follows: (i) item response model, (ii) item pool, (iii) entry level, (iv) item selection rule, (v) scoring method (ability estimation method), (vi) termination criterion. Thompson (2007), on the other hand, posits that although CCT applications are similar to CAT procedures, they consists of five basic technical components: (i) psychometric model, (ii) calibrated item bank, (iii) starting point, (iv) item selection algorithm, (v) termination criterion (classification/scoring procedure). Therefore, it is possible to suggest that CCT applications differ from CAT in terms of ability estimation. In this sense, Weiss and Kingsbury (1984) suggest that the test administration should be terminated when the ability estimation of examinee is achieved with the desired precision in CAT, In this sense, Weiss and Kingsbury (1984) suggest that the test administration should be terminated when the expected ability estimation of examinee is achieved in CAT, but according to Thompson (2007), it should be terminated when the examinee has been classified in one of the predetermined categories.

In addition to the components mentioned above related to individualized tests, such constraints as item exposure control and content balancing in CAT applications can also be considered in CCT applications in order to achieve valid, reliable measurements with high level of test security. Although content balancing and item exposure control are simultaneously regarded as additional constraints in recent CAT studies, these restrictions are not taken into account in many CCT studies. In this regard, this gap in the literature is worth to examine the issues about content balancing and item exposure control.

The purpose of this research was to evaluate the effect of item pool and selection algorithms (e.g. item selection methods, content balancing methods, and item exposure control methods) on CCT performance in terms of classification evaluation metrics. Within the frame of this purpose, CCT simulations were performed on different size item pools, consisting of unidimensional and dichotomous items and showing peaked and broad distributions. In the evaluation of CCT performance, average classification accuracy (ACA), average test length (ATL), applied content rates, the proportion of overexposed items in the pool (i.e., the proportion of items in the pool with exposure rate exceeding r_{max}) (OEX), the mean exposure rate of overexposed items (MOEX), and test overlap values were used in the current study.

The sub-problems of the current study were identified as follows:

How do ACA, ATL, applied content rates, OEX, MOEX and test overlap values in the item pools with 150, 300, 450 and 600 change in the cases where the item pools show peaked or broad distributions based on dichotomous classifications in which the sequential probability ratio test (SPRT) classification criterion is employed together with cutscore based item selection methods related to the Maximum Fisher Information (MFI-CB) and the Kullback-Leibler Information (KLI-CB), and content balancing methods such as constrained CAT (CCAT) and modified multinomial model (MMM), and item exposure control methods including Sympson-Hetter (SH) and item eligibility (IE)?

The previous literature has shown a wide range of studies on applications. Among these are classification criteria (e.g., Kingsbury & Weiss, 1980; Spray & Reckase, 1996; Thompson, 2009), item selection methods (e.g., Eggen, 1999; Lin & Spray, 2000) or classification criteria crossing with different item selection methods (e.g., Eggen & Straetmans, 2000; Thompson & Ro, 2007). There have been, on the other hand, few studies on such constraints as content balancing and/or item exposure control. Even though there are concerns over using these constraints on the research conditions may eliminate the differences between item selection methods, it is an undeniable fact that content balancing is required for tests covering content areas based on maximum information and with high content validity, while item exposure control is for test security that can be increased through the proper utilization of the item pools (Leroux et al., 2019; Lin,

2011). Given the advantages by tests with high validity and test security to the classification accuracy of the examinees, it is possible to consider new CCT applications in which content balancing and item exposure control are considered in the research constraints as a new contribution to the literature. In this sense, this study is thought to useful for making significant contributions to the literature. For the purpose of the research, below are described design of the research, data generation, CCT simulation conditions and data analysis process. Then, the findings are presented and the results of the research are discussed. Finally, recommendations are given for future applications and research.

2. Methodology

This study is a descriptive one examining item selection methods, content balancing methods, and item exposure control methods used in CCT applications over different size item pools with peaked and broad distributions. Descriptive studies are those in which a given state of affairs is described as thoroughly as possible (Fraenkel, Wallen & Hyun, 2012). This study is a simulation research, as well. Simulation studies, which form and analyse many different research conditions simultaneously, allow researchers to examine more complex research designs (Dooley, 2002). The dependent variables in this study are ACA, ATL, applied content rates, OEX, MOEX, and test overlap rates, whereas the independent variables are item selection methods (MFI-CB, KLI-CB), content balancing methods (CCAT, MMM), item exposure control methods (SH, IE), item pool distribution (peaked, broad) and item pool sizes (150 items, 300 items, 450 items, 600 items).

2.1. Data Generation

In this study, the three-parameter logistic model from the IRT models was preferred and the item pools were designed such a way that they had 150, 300, 450 and 600 items having peaked and broad distributions. The three-parameter logistic model was used in that it would contribute to the classification accuracy by taking into account the guessing parameter (parameter *c*) as well as the item discrimination parameter (parameter *a*) and the item difficulty parameter (parameter *b*). For the items in the pools, the item discrimination parameter *a* was generated between $a \sim U[0.5, 2.0]$ range as it had uniform distribution for medium and high discrimination based on the study conducted by Kingsbury and Weiss (1980). The item difficulty parameter *b*, on the other hand, was generated as $b \sim N(0, 0.4)$ in the item pools having a peaked distribution and as $b \sim N(0, 1.5)$ in the item pools with a broad distribution for getting close to the real values considering the study of Thompson (2009). Finally, the guessing parameter *c* was generated as $c \sim N(0.20, 0.05)$ from the normal distribution by considering the probability of selecting the correct item option by 20%. Further, ability parameters were randomly generated in the R program in a way that 1000 examinees with normal distribution were $\Theta \sim N(0, 1)$. In addition, each examinee's item response pattern was also simulated in the R program.

2.2. CCT Simulation Conditions

Below are some explanations about CCT simulation conditions as sub-headings.

2.2.1. Starting point

In CCT applications, examinee ability at the mean of the population or a likelihood ratio of 1.0 (an even ratio) is often used as the starting point (Thompson, 2007). Based on this, the starting point $\theta = 0$ was determined as the default value in all conditions in the current study.

2.2.2. Item selection

Intelligent item selection methods, based on the random selection of the best item among the unused items in the pool, are often classified into two types, cutscore based and estimated based (Thompson, 2007). The cutscore based (CB) methods related to the IRT are Maximum Fisher information (MFI), Maximum Kullback-Leibler information (KLI) and log-odds ratio criterion. Accordingly, these three methods are comparable with each other (Lin & Spray, 2000).

The MFI method aims to maximize the information at a cut point (θ), the probability of a correct answer *P* and the probability of a wrong answer *Q*, and is formulated by the following equation (Embretson & Reise, 2000).

$$I_i(\theta) = [\frac{\partial P_i(\theta)}{\partial \theta}]^2 / P_i(\theta) Q_i(\theta)$$

The KLI method, on the other hand, evaluates the information between θ_0 and θ_1 around the nearest cutting point as the probability of a correct answer *P* and the probability of a wrong answer *Q* and the equation below displays the sum of KLI of the items (Eggen, 1999).

$$K_i(\theta_1 \| \theta_0) = P_i(\theta_1) \log \frac{P_i(\theta_1)}{P_i(\theta_0)} + Q_i(\theta_1) log \frac{Q_i(\theta_1)}{Q_i(\theta_0)}$$

In this current study, MFI-CB and KLI-CB from cutscore based item selection methods were used.

2.2.3. Ability estimation

Maximum Likelihood Estimation (MLE), Marginal Maximum Likelihood Estimation (MMLE), Weighted Likelihood Estimation (WLE), Maximum A Posteriori (MAP), and Expected a Posteriori (EAP) are the most used ability estimation methods from the *unidimensional IRT* model applied in the literature. These methods, however, are subjected to several disadvantages. Accordingly, Warm (1989) suggests that all these estimation methods can produce biased estimates to some degree. Bias is a major disadvantage for CCT applications in that it can systematically affect the precision of the cut score (Wang & Wang, 2001). When the item pool size is small, the EAP may bias results toward the mean of estimated ability levels. The previous study has indicated that increasing the number of items reduces the bias of maximum likelihood estimates, but there is no clear-cut answer about how many items will reduce the bias (Wainer & Thissen, 1987). In this sense, considering the item pool sizes in this current study, the EAP ability estimator method which can make unbiased estimations as much as possible, was used.

2.2.4. Classification criteria

There are three basic classification criteria based on IRT in CCT 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). Further, Thompson (2009) notes that while the item selection method is based on the cutscore in both peaked and broad distributions of item pools, the SPRT classification criterion is more useful in terms of test efficiency. Accordingly, in the present research, based on the cutscore based item selection, the SPRT classification criterion, frequently used in the literature, was preferred. In addition, considering the research conditions for SPRT, the indifference region constant was determined as δ: .35.

2.2.5. Content balancing

An examination of previous research has shown that content balancing methods often used in CCT applications are the spiraling method (Kingsbury & Zara, 1989) (e.g., Finkelman, 2008; Huebner, 2012) and the constrained CAT (CCAT) method (e.g., Eggen & Straetmans, 2000; Huebner & Li, 2012). Unlike

the previous research in the literature, in this study, MMM method, used by Lin (2011), was employed so as to determine performance differences among methods in addition to CCAT method.

While a test is performed, the content balancing process via CCAT method is as follows (Kingsbury & Zara, 1989).

1st Step: Following the administration of an item, examinees' provisional ability level estimate is calculated.

2nd Step: Percentage of items which have already been administered in each content area are calculated.

3rd Step: Calculated percentages and predetermined desired percentages are compared and content area with the largest discrepancy is identified.

4th Step: The most informative item at the examinees' provisional ability level estimate from the content area with the largest discrepancy are selected and administered it to the examinees.

These steps are iteratively performed after each item until the test is terminated. Then, the test is completed when any predetermined test termination criterion is satisfied.

In MMM method, on the other hand, following a multinomial distribution, a cumulative distribution is created based on prespecified content rates for each sub-content area. Then, a number is randomly selected from a uniform distribution ranging from 0 to 1, and the most appropriate item selected from the sub-content area to which this number corresponds is administered to the examinees. For example, in this current study, in which the desired content coverages were 0,45, 0,35 and 0,20 respectively, the cumulative percentage for contents 0,45, 0,80 (0,45 + 0,35) and 1 (0,80 + 0,20) were respectively. If 0,3 is selected as a random number in a uniform distribution ranging between 0 and 1, the first sub-content is selected. If a random number 0.7 is chosen, the second sub-content is selected The iterative process is carried out through the administration of the best item in the chosen sub-content to the examinee. A randomized content area sequence prevents content sequencing predictability. The randomized content selection ends when a desired content percentage is achieved (Lin, 2011).

In this study, the minimum number of items to be used before terminating the test was limited to 5 and the maximum number of items to 30, considering the research conditions such as item pool sizes, content balancing and item exposure control. Item pools designed with 150, 300, 450, and 600 items derived in the R program when included under the content balancing research conditions with CCAT or MMM were divided into three content areas with the random item assignment. Then, with the help of loops written by the researcher, item selection was performed based on these content areas. The desired percentages of the prespecified content areas were adjusted to be 45%, 35% and 20%, respectively.

2.2.6. Item exposure control

The random item selection from randomized strategies and SH method from conditional selection strategies are the most widely established item exposure control procedures (Sympson & Hetter, 1985). In the present study, SH and IE methods which are considered as more effective than random item selection and are based on the conditional selection strategy under realistic test conditions (van der Linden & Veldkamp, 2004) were used. SH and IE methods check the desired maximum item exposure rate (r_{max}), by assigning an exposure control parameter (K_m) to each item in the pool. The difference between the methods is how and when these K_m parameters are calculated (Huebner, 2012). In SH method, K_m is a constant parameter and is calculated through iterative simulations before the administration of the test to examinees as follow:

$$K_m = \begin{cases} 1 & if \quad P(S_m) \leq r_{max} \\ \dfrac{r_{max}}{P(S_m)} & if \quad P(S_m) > r_{max} \end{cases}$$

In each iteration of these preliminary simulations, the probability $P(S_m)$ of choosing m items is recalculated and the item exposure control parameters are updated consistent with this rule (Huebner, 2012).

IE method, on the other hand, needs no preliminary simulations and $K_m(i+1)$ is updated using the following rule at the time of administration of the test, with the item exposure control parameter based on individuals *1,2, ..., i* and the probability of administration of the item m $P(A_m)$ (Huebner, 2012).

$$K_m{}^{(i+1)} = \begin{cases} 1 & if \quad \dfrac{P^{(1\ldots i)}(A_m)}{K_m{}^{(i)}} \leq r_{max} \\ \dfrac{r_{max} K_m{}^{(i)}}{p^{(1\ldots i)}(A_m)} & if \quad \dfrac{P^{(1\ldots i)}(A_m)}{K_m{}^{(i)}} > r_{max} \end{cases}$$

In this study, the desired maximum item exposure rate was defined as r_{max}=.20, which is thought to be an average rate (Huebner, 2012; Leung, Chang & Hau, 2002; Thompson & Ro, 2007).

2.2.7. Classification categories and cutscore

In this present study, in which dichotomous classifications were made, the cutscore was set according to the generated ability parameters of examinees. Similarly, in the study by Eggen and Straetmans (2000), the first half of the ability parameters ranked from low to high were determined as level 1 and the second half as level 2. Then 70% of the highest ability in the level 1 was taken as the cutscore (CS = 0.00).

2.2.8. Item pool sizes and distributions

The simulations were designed to examine the effect of item pool distribution and item pool size in the CCT application. Considering the study of Thompson (2009), a total of eight item pools (item pools of 150, 300, 450 and 600 items with a peaked distribution and item pools of 150, 300, 450 and 600 items with a broad distribution) in the R program were designed.

2.3. Data Analysis

In a CCT application, it is mostly aimed at high ACA, low ATL, OEX, MOEX, test overlap rates and as much as possible along with the applied content rates providing the desired content rates. As a result of research by Harwell et al. (1996), a minimum of 25 replications were proposed for MC studies in IRT-based research. In this regard, for the evaluations for these goals, 100 iterations were carried out for each of the 64 simulation conditions and the values of the dependent variables were obtained by calculating the average of the iterations. The test overlap rate was calculated by the following formula employed by Huo (2009).

$$\frac{(\sum C_0)/C_N^2}{(\sum_{i=1}^{N} J_i)/N}$$

In this formula, N is the number of examinees, C_0 is the number of common items for any two examinees, $C_N{}^2$ is the total number of possible pairs of N examinees, and J_i is the test length of examinee i.

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

3. Findings

Findings obtained in the study are presented in this section.

Table 1 shows the values calculated by averaging 100 replications performed for each simulation condition related to the peaked distribution of item pools.

Table 1. Comparison of the CCT applications over the peaked distribution of ıtem pools to different sizes.

IPS	ISM	CBM	IECM	ACA	ATL	Applied Content Rates			OEX	MOEX	Test Overlap
150	MFI-CB	CCAT	SH	.91	11.79	46.79	34.16	19.05	.20	.34	.32
			IE	.90	13.35	45.85	34.28	19.87	.17	.21	.19
		MMM	SH	.91	11.80	45.10	35.06	19.85	.20	.33	.31
			IE	.90	13.59	45.16	35.14	19.70	.17	.20	.18
	KLI-CB	CCAT	SH	.91	11.82	46.80	34.16	19.04	.20	.34	.32
			IE	.90	13.63	45.79	34.27	19.94	.17	.21	.19
		MMM	SH	.91	11.77	44.97	35.07	19.96	.20	.33	.31
			IE	.90	13.57	45.02	35.04	19.94	.18	.21	.19
300	MFI-CB	CCAT	SH	.91	10.84	47.48	34.06	18.46	.09	.34	.32
			IE	.91	11.66	46.91	34.17	18.92	.07	.21	.19
		MMM	SH	.91	10.80	45.07	34.92	20.00	.09	.33	.31
			IE	.91	11.57	44.98	35.05	19.97	.07	.21	.18
	KLI-CB	CCAT	SH	.91	10.83	47.51	34.10	18.39	.09	.34	.32
			IE	.91	11.68	46.96	34.14	18.90	.07	.21	.19
		MMM	SH	.91	10.81	45.03	35.05	19.91	.09	.33	.31
			IE	.91	11.63	45.05	34.98	19.97	.07	.21	.18
450	MFI-CB	CCAT	SH	.91	10.50	47.78	34.06	18.16	.06	.34	.32
			IE	.91	11.04	47.45	34.07	18.48	.05	.21	.18
		MMM	SH	.92	10.44	45.05	34.97	19.99	.06	.33	.31
			IE	.91	11.04	45.03	35.04	19.94	.04	.21	.18
	KLI-CB	CCAT	SH	.92	10.43	47.83	34.03	18.13	.06	.34	.32
			IE	.91	11.07	47.40	34.06	18.54	.05	.21	.18
		MMM	SH	.91	10.44	45.00	35.00	20.00	.06	.33	.31
			IE	.91	11.06	44.98	34.99	20.03	.04	.21	.18
600	MFI-CB	CCAT	SH	.92	10.18	48.10	33.96	17.94	.04	.34	.31
			IE	.91	10.74	47.58	34.09	18.33	.03	.21	.18
		MMM	SH	.92	10.20	44.95	35.03	20.02	.04	.33	.31
			IE	.91	10.74	44.89	35.03	20.08	.03	.21	.18
	KLI-CB	CCAT	SH	.91	10.28	47.94	34.01	18.05	.04	.34	.32
			IE	.91	10.75	47.63	34.06	18.30	.03	.21	.18
		MMM	SH	.92	10.25	45.01	35.02	19.97	.04	.33	.31
			IE	.91	10.73	45.03	35.03	19.94	.03	.21	.18

Note: IPS= item pool size, ISM= item selection method, CBM= content balancing method, IECM= item exposure control method, ACA= average classification accuracy, ATL= average test length, OEX= the proportion of overexposed items in the pool, MOEX= the mean exposure rate of overexposed items, MFI-CB= maximum fisher information method based on cutscore, KLI-CB= Kullback-Leibler information method based on cutscore, CCAT= constrained computerized adaptive testing, MMM= modified multinomial model, SH= Sympson-Hetter method, IE= item eligibility method.

As can be seen from the table 1, in item pools with a peaked distribution, high classification accuracy (between 90% and 92%) was obtained under all research conditions and item pool size did not affect ACA. In terms of ATL, on the other hand, there were similar results for both item selection methods in item pools of the same size. The results showed that there was a slight decrease in ATL as the item pool size increased. Furthermore, especially in the 150-item item pool, regardless of the item selection method and the content balancing method, when SH item exposure control method was used, there was slightly higher ACA but lower ATL compared to IE method. Accordingly, it can be noted that test efficiency is higher in these conditions. However, it is seen that this difference decreases, even almost disappears as the item pool size increases. When MMM was employed as the content balancing method, the applied content ratios provided the desired content ratios (45%, 35% and 20%, respectively). On the other hand, when CCAT method was used, the applied content rates were partially above or below the desired content rates. There is evidence to suggest that OEX rates are higher in the 150-item pool compared to other item pools. On the other hand, it

is seen that the MOEX rates are similar for the same conditions in all item pools. For instance, in the 150-item pool, under the conditions in which MFI-CB, CCAT and SH methods were used together, approximately 20% of the items exceeded the item exposure rate (r_{max} = .20), and the mean exposure rate of overexposed items was calculated as approximately .34. Further, it is important that OEX, MOEX and test overlap rates were calculated lower regardless of other research conditions when IE item exposure control method was used.

Table 2 shows the values calculated by averaging 100 replications performed for each simulation condition related to the broad distribution of item pools.

Table 2. Comparison of the CCT applications over the broad distribution of ıtem pools to different sizes.

IPS	ISM	CBM	IECM	ACA	ATL	Applied Content Rates			OEX	MOEX	Test Overlap
150	MFI-CB	CCAT	SH	.90	14.28	45.35	34.34	20.31	.25	.36	.33
			IE	.87	18.16	43.99	34.40	21.60	.27	.21	.19
		MMM	SH	.89	14.35	44.99	35.02	20.00	.25	.33	.32
			IE	.88	17.23	45.10	34.88	20.02	.21	.21	.19
	KLI-CB	CCAT	SH	.89	14.43	45.30	34.33	20.37	.25	.33	.33
			IE	.87	17.09	44.26	34.39	21.34	.22	.21	.19
		MMM	SH	.89	14.43	45.04	34.97	19.98	.25	.33	.32
			IE	.89	16.84	45.06	34.89	20.05	.19	.21	.19
300	MFI-CB	CCAT	SH	.91	12.42	46.41	34.24	19.35	.11	.34	.32
			IE	.90	14.34	45.34	34.34	20.32	.09	.21	.19
		MMM	SH	.90	12.34	45.01	34.99	20.00	.10	.33	.31
			IE	.90	14.10	44.95	35.05	20.00	.09	.21	.19
	KLI-CB	CCAT	SH	.91	12.44	46.40	34.21	19.39	.11	.34	.32
			IE	.90	14.13	45.48	34.29	20.24	.09	.21	.19
		MMM	SH	.91	12.46	45.03	34.97	20.00	.11	.33	.31
			IE	.90	14.23	44.97	35.06	19.96	.09	.21	.19
450	MFI-CB	CCAT	SH	.91	11.68	46.90	34.16	18.94	.07	.34	.32
			IE	.91	13.01	46.11	34.22	19.67	.05	.21	.19
		MMM	SH	.91	11.74	44.98	35.00	20.01	.07	.33	.31
			IE	.90	12.99	45.06	34.96	19.98	.05	.21	.19
	KLI-CB	CCAT	SH	.91	11.72	46.85	34.20	18.95	.07	.34	.32
			IE	.90	12.93	46.09	34.26	19.65	.05	.21	.19
		MMM	SH	.91	11.65	45.04	34.94	20.02	.07	.33	.31
			IE	.91	12.92	45.02	34.99	20.00	.05	.21	.19
600	MFI-CB	CCAT	SH	.91	11.30	47.06	34.18	18.76	.05	.34	.32
			IE	.91	12.35	46.45	34.22	19.33	.04	.21	.19
		MMM	SH	.91	11.30	45.01	34.95	20.04	.05	.33	.31
			IE	.91	12.33	44.91	35.03	20.05	.04	.21	.18
	KLI-CB	CCAT	SH	.91	11.27	47.25	34.07	18.68	.05	.34	.32
			IE	.91	12.36	46.40	34.22	19.38	.04	.21	.19
		MMM	SH	.91	11.23	45.00	35.03	19.97	.05	.33	.31
			IE	.91	12.24	44.92	35.05	20.03	.04	.21	.18

Note: IPS= item pool size, ISM= item selection method, CBM= content balancing method, IECM= item exposure control method, ACA= average classification accuracy, ATL= average test length, OEX= the proportion of overexposed items in the pool, MOEX= the mean exposure rate of overexposed items, MFI-CB= maximum fisher information method based on cutscore, KLI-CB= Kullback-Leibler information method based on cutscore, CCAT= constrained computerized adaptive testing, MMM= modified multinomial model, SH= Sympson-Hetter method, IE= item eligibility method.

As can be seen from the Table 2, in the item pools with a broad distribution, high classification accuracy (range 87% to 91%) was obtained under all research conditions, and when compared to Table 1, it is seen that ACA decreased especially in 150-item pools with a broad distribution compared to item pools with a peaked distribution, but ACA did not change much in larger size item pools. When Table 2 is compared to Table 1 in terms of ATL, there needed more items, particularly in the item pool with 150-items, to terminate

the test, and there was a slight decrease in ATL as the item pool size increased. In addition, there were similar results in terms of ACA and ATL in the conditions created for both item selection methods used in item pools with the same sizes. According to Table 2, it is seen that the OEX rates are higher in the 150-item pool compared to other item pools, and when compared to Table 1, the OEX rates in the item pools with a broad distribution are higher. Furthermore, in line with the findings on Table 1 regardless of item selection method, test efficiency is higher when SH item exposure control method is used, but this difference between SH and IE decreases as the item pool size increases. MMM method outperforms CCAT method in providing the desired content rates, and the MOEX values are similar in all item pools under the same conditions. Further, it was concluded that OEX, MOEX and test overlap rates were higher in general when SH item exposure control method was used.

4. Conclusion and Suggestions

In this study, the comparison of cutscore based item selection methods, content balancing methods, and item exposure control methods used in CCT applications in terms of different size item pools with peaked and broad distributions were examined.

The findings of the study revealed that high classification accuracy was obtained in all research conditions in the item pools with both peaked and broad distribution. Moreover, it was determined that ACA was partially lower, while ATL was higher in item pools with a broad distribution. This difference observed between peaked and broad distributions of item pools in terms of ACA and ATL is more evident especially in the smallest item pool (150-item pool). This finding corraborated with the study by Thompson (2009). Accordingly, Thompson (2009), concluded that while ACA had similar values, ATL was higher in broad item pools in the comparison of item pools with peaked and broad distributions. Another finding in line with Thompson (2009) is that in both peaked and broad distributions of item pools, there was a slight decrease in ATL values as the item pool size increased. In addition, it was found out that ACA increased along with the decrease in ATL in the item pools larger than 150 items. The item selection methods showed similar results in terms of ACA and ATL. The similar results by the item selection methods MFI and KLI may have derived from the fact that these two methods were similar in their nature. The items with maximum information at the examinee's recent ability is selected in the ability estimation through MFI, whereas the items with maximum information is preferred at the bounds of the indifference region in the ability estimation through KLI (Spray & Reckase, 1994). Regardless of the distribution of the item pool, it was concluded that in the conditions where SH item exposure method was used, slightly higher ACA and lower ATL were calculated compared to IE method. Accordingly, it is seen that SH method is more advantageous in terms of test efficiency. However, this difference by SH and IE decreased as the item pool size increased. In addition, OEX, MOEX and test overlap values were calculated lower when item exposure control was performed with IE method. Accordingly, it can be noted that IE method is more advantageous in terms of item exposure control. Huebner (2012) supported this finding by concluding that IE method is more effective than SH method in terms of item exposure control. Based on this finding, it can be concluded that IE method needs as large item pools as possible in order to perform better in terms of test efficiency as well as item exposure control. There is also evidence that content balancing with MMM method provided desired content rates in all conditions, and therefore performed better than CCAT. Similarly, Lin (2011) found that compared to the most frequently used content balancing methods in CCTs, MMM method, mostly used in CATs in the literature, is more successful in providing the desired content balance as it successfull controlled content balancing. The results of this study indicated that OEX rates were higher in the item pools with 150 items and in those with a broad distribution. On the other hand, MOEX and test overlap rates were similar under the same conditions in item pools with both peaked and broad distributions. The reason why the test overlap values did not change despite the larger size of the item pool may be the fact that the item pools show similar conditions in terms of test overlap and that they include at least 150 items which make them large enough. Accordingly, based on the results obtained from the current study, it

can be noted that it is worth to use item pools larger than 150 items as much as possible to increase the ACA value and decrease the ATL and OEX values in line with the expectations in CCT applications. Furthermore, for more effective CCT applications in which item exposure control and content balancing are performed, the use of larger item pools should be preferred when the item pools with broad distributions are used compared to those with peaked distributions.

Considering the results of this paper, there are some practical implications. In order to perform CCT applications with higher ACA and lower ATL (with higher test efficiency), it can be recommended that the item pool should be as close to a peaked distribution as possible and SH method should be preferred. On the other hand, if item exposure control is of critical importance, the more advantageous IE method can be employed. Further, the use of MMM method may be preferred for content balancing purposes. In general, as large item pools as possible can be used to gain advantages in terms of test efficiency, item exposure control, and content balancing. In this regard, as large item pools as possible can be used to be able to provide advantage in terms of test efficiency, item exposure control, and content balancing. In future studies, comparing the item pools with peaked, broad, and normal distributions in terms of different classification criteria, ability estimation methods, and item selection methods can provide contribution to the related literature. In addition, these comparisons can be conducted by considering multi-dimensional item pool or real data sets.

References

Bao, Y., Shen, Y., Wang, S., & Bradshaw, L. (2020). Flexible computerized adaptive tests to detect misconceptions and estimate ability simultaneously. *Applied Psychological Measurement, 45*(1), 3-21. https://doi.org/10.1177/0146621620965730

Dooley, K. (2002). Simulation research methods. In J. Baum (Ed.), *Companion to organizations* (pp. 829-848). 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. https://doi.org/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. https://doi.org/10.1177/00131640021970862

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

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. https://doi.org/10.3102/1076998611422912

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

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. https://doi.org/10.3102/1076998607308573

Hambleton, R. K., & Swaminathan, H. (1985). I*tem response theory: Principles and applications*. Kluwer Nijhoff Publishing.

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. https://doi.org/10.1177/014662169602000201

Huebner, A. (2012). Item overexposure in computerized classification tests using sequential item selection. *Practical Assessment, Research & Evaluation, 17*(12), 1-9. https://doi.org/10.7275/nrlc-yv82

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

Huo, Y. (2009). *Variable-length computerized adaptive testing: adaptation of the a-stratified strategy in item selection with content balancing*. Unpublished doctoral dissertation. University of Illinois, Champaign. http://hdl.handle.net/2142/14715

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. 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). 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. https://doi.org/10.1207/s15324818ame0204_6

Krabbe, P. F. M. (2017). The Measurement of Health and Health Status: Concepts, Methods and Applications from a Multidisciplinary Perspective. *New Developments* (Ch.14, ss. 309-331). Academic Press. https://doi.org/10.1016/B978-0-12-801504-9.00014-3

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. https://doi.org/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. https://doi.org/10.1177/014662102237795

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. *Applied Psychological Measurement 71*(1), 20-36. https://doi.org/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. https://eric.ed.gov/?id=ED445066

Spray, J. A. & Reckase, M. D. (1994). The Selection of Test Items for Decision Making with a Computer Adaptive Test. *The Annual Meeting of the National Council on Measurement in Education*. NewOrleans, LA, 5-7 April 1994. https://eric.ed.gov/?id=ED372078

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. https://doi.org/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. http://www.iacat.org/content/controlling-item-exposure-rates-computerized-adaptive-testing

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment, Research & Evaluation, 12*(1), 1-13. 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. https://doi.org/10.1177/0013164408324460

Thompson, N. A. (2011). Termination criteria for computerized classification testing. *Practical Assessment, Research & Evaluation, 16*(4), 1-7. https://doi.org/10.7275/wq8m-zk25

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*. 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. https://doi.org/10.3102/10769986029003273

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*(4), 339–368. https://doi.org/10.2307/1165054

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://doi.org/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

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