ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 395-407

Effect Of Content Balancing on Measurement Precision in Computer Adaptive Testing Applications

İlkay ÜÇGÜL ÖCAL *   Nuri DOĞAN **

Abstract

This study aims to investigate the effect of content balancing, which involves equal and different weighting of content areas in dichotomous items in computerized adaptive testing (CAT), on measurement precision under different measurement conditions. Conducted as a simulation study, small sample sizes were set at 250, while large sample sizes comprised 500 individuals. The ability parameters of the individuals forming the sample were generated to display a normal distribution within the range of -3 to +3 for each sample. Using the three-parameter logistic (3PL) item response model, a pool of 750 items spanning five different content areas was developed for dichotomous items. The study considered different sample sizes, ability estimation methods (Maximum Likelihood Estimation and Expected A Posteriori), and termination rules (20 items, 60 items, and SE≤.30) as significant factors in the CAT algorithm for examining the effect of content balancing. For each CAT application, measurement precision was assessed by calculating the root mean square error (RMSE), bias, and fidelity coefficients, and these were analyzed comparatively. The results showed that bias values were close to zero under all conditions. RMSE values were lowest when the test was terminated at 60 items across all conditions, while standard error termination rules and situations where the test terminated at 20 items produced similar values. Considering all conditions, the highest fidelity coefficient was observed when the test terminated at 60 items. The fidelity coefficient did not vary significantly with other variables. Implementing content balancing in conditions using different ability estimation methods increased the average number of items by approximately one item. While the average number of items in the test slightly increased with content balancing, measurement precision was maintained. Overall, the maximum item exposure rate decreased with content balancing when content areas were weighted equally, whereas it increased when they were weighted disproportionately.

Keywords: *computerized adaptive testing, content balancing, measurement precision.*

Introduction

Examinations used in education have traditionally focused on paper and pencil tests and performance assessments. Since the late 1980s, with the widespread adoption of personal computers in education, these examinations have rapidly expanded into formats suitable for computer delivery (Şenel, 2021; Van der Linden & Glas, 2002). Computerized adaptive tests (CATs) utilise an algorithmic approach to administer test items. Specifically, the items selected and administered are tailored to the estimated ability level of the examinee during the testing process, with the estimated ability continually updated after each item is administered. Therefore, CAT is an adaptable test at the item level and can be of fixed or variable length. Ability estimation is used not only to represent an examinee's level of ability but also to determine the selection of subsequent items from the available item pool. CATs can be considerably more useful and efficient than traditional linear tests, which has led to their widespread use in recent years (Cheng & Chang, 2007; Kalender, 2009). Several advantages of CATs over traditional linear tests have been demonstrated, including increased flexibility in test administration, elimination of the need for answer sheets and trained test administrators, enhanced test security, and the ability to provide accurate measurements across a wide range of ability levels (Rudner, 1998; Tian et al., 2007).

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications, *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

The mathematical model used for CAT applications is based on Item Response Theory (IRT). IRT methodologies are employed in various CAT processes and focus on improving the accuracy and efficiency of ability estimation. IRT-based CAT applications typically contain fewer items than traditional paper and pencil measurements (Embretson & Reise, 2000). The CAT process requires a calibrated item pool and is implemented in four consecutive steps (Thompson & Weiss, 2019):

1. The initial step involves selecting one or several items to start the CAT.
2. The testing step, where items are selected iteratively and optimally, is administered, and ability estimation is performed after each item administration.
3. The termination step defines rules for stopping the adaptive item administration.
4. The final step involves final ability estimation and reporting.

The initial step involves selecting the first item(s) to be administered in the CAT. A commonly used starting rule is the selection of an item that corresponds to the average ability level of the examinee group (theta=0). If no information is available about the examinees' ability levels at the start, this method is considered appropriate. An alternative entry rule could be the selection of an item of medium difficulty ($-0.5<b<0.5$) at the start. After administering the initial item, the cycle of ability estimation and item administration continues until the testing process concludes. Various estimation methods are available for ability estimation. The most commonly used methods include Maximum Likelihood Estimation (MLE), Maximum A Posteriori (MAP), and Expected A Posteriori (EAP). When item parameters are known, ability parameters can be simply estimated using the maximum likelihood estimation method. This method has several advantages, including consistency and asymptotic normality. For the MLE estimation method to be applicable, the response pattern must contain at least one correct and one incorrect answer. In cases where all items are answered correctly or all are answered incorrectly, the use of the MLE estimation method is not appropriate. In such cases, Bayes-based ability estimation methods, such as EAP or MAP, can be used to overcome this problem. Bayes-based estimation methods have smaller standard errors compared to MLE but require prior knowledge of the individual's ability. The choice of which ability estimation method to use should be made considering all components of the CAT application (Hambleton & Swaminathan, 1985). In the testing step, a hybrid rule that starts with one estimation method and then switches to another after a certain number of items or under certain conditions can also be preferred (Magis et al., 2017). Item selection is a critical component of CAT applications. After determining that test items are appropriate based on the content characteristics in the content balancing component of the CAT algorithm, these items are considered for selection as the next item to be administered. A comprehensive range of item selection methods has been developed in the testing measurement field, yet very few of these methods are employed in actual CAT applications (Han, 2018). One of the best-known and oldest item selection methods is the Maximum Fisher Information (MFI) method. This method involves selecting an item that has the MFI at a certain θ based on the test items previously administered to the examinee.

Test developers have found that the choice of termination rule is largely dependent on the test purpose, item pool characteristics, and operational constraints (Segall, 2005). The termination rule defines parameters for stopping the adaptive item administration. In general, four main termination rules are identified: (a) length criterion, (b) precision criterion, (c) classification criterion, and (d) information criterion (Van der Linden & Glas, 2002).

Validity is one of the most crucial characteristics sought in tests used in education and psychology. Validity refers to "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA & NCME, 2014). The constructs measured in educational tests are combinations of different subject and content areas. Ensuring the content validity of a test, representing these subjects and content areas adequately within the test is possible. Depending on the test requirements, item selection in CAT applications must meet the requirements of the defined scope to have a balanced content representation; that is, the CAT application must balance items from each subject area according to predetermined percentages. In individualised tests, different items are administered to examinees, but the same item distribution according to the content area should be provided to each. To obtain valid measurements, there must be a balance between the measured content areas or subject areas. Several content balancing methods have been developed to ensure that CAT maintains the desired distribution of content areas throughout the test. Among the most widely used methods are the Constrained Computerized Adaptive Testing (CCAT), the Modified Multinomial Model (MMM), and the Modified CCAT (MCCAT).

The CCAT method, proposed by Kingsbury and Zara (1989), is a straightforward and understandable two-stage content balancing control mechanism. The content balancing algorithm selects the most suitable item from a content area with the current item usage frequency rate below the targeted application percentage. The selection of the most suitable item is limited by the item usage frequency rate and content area determined to be below the target percentage for the test. Content areas can be weighted equally or differently according to the structure of the respective course. In this method, at each step of the CAT process, experimental percentage rates for each category are calculated. Subsequently, the category with the greatest difference between the theoretical and experimental values is identified, and the next item is selected from this subgroup before returning to the first step. Based on this method, any desired content distribution can be met if the number of items in each content area in the item pool is sufficiently large to construct the target test. The MMM, as described by Chen and Ankenmann (2004), begins by constructing a cumulative distribution based on the target exposure rates of all content areas. A random number from a uniform distribution is used to select the next content area. When a content area reaches its target percentage, a new multinomial distribution is created using the remaining content areas. This method avoids the highly predictable sequence of content areas seen in the CCAT and ensures that target percentages are met exactly. The MCCAT method, proposed by Leung et al. (2000), modifies the original CCAT by selecting items from any unfulfilled content area rather than the one furthest below its target. This approach helps avoid potential undesirable order effects of the CCAT, ensuring a more balanced and less predictable item selection process.

Decisions made based on the measurement results obtained from CAT applications have significant impacts on all educational stakeholders. Therefore, it is crucial to make valid and reliable estimations with CAT applications. The lack of content comparability can pose a threat to the content validity of scores. Whether or not to balance the content of items administered to examinees is one of the fundamental issues to be addressed when developing a CAT application.

Previous studies have extensively explored various aspects of content balancing in CAT. Cheng and Chang (2007) investigated a two-phase item selection procedure that adapts to content requirements while optimizing item selection, highlighting the impact of flexible content balancing on measurement precision and efficiency. Leung et al. (2000) introduced the MCCAT method, which eliminates the predictability of content sequencing while maintaining balance. In subsequent studies, Leung et al. (2003a, 2003b) examined the multistage a-stratified design (ASTR) combined with content balancing methods like MCCAT and the MMM, demonstrating the effectiveness of these methods in reducing item-overlap rates and enhancing item pool utilization without compromising measurement accuracy. Furthermore, Özdemir and Gelbal (2015) and Sari and Manley (2017) explored the practical applications of content balancing in educational settings, emphasizing its role in maintaining test reliability and validity. Demir (2019) analyzed the effects of content balancing on the precision and fairness of CAT applications, providing insights into the psychometric properties affected by different balancing algorithms. Şahin and Özbaşı (2017) reviewed various content balancing methods, offering a comprehensive overview of the current state of research and practical implications. Additionally, Song (2010) focused on the implementation challenges and solutions for content balancing in large-scale adaptive testing programs, while Yasuda and Hull (2021) demonstrated the application of content balancing in the development of CAT-based versions of specific inventories, showing that it can be implemented without compromising accuracy. However, these aforementioned studies often focused on specific methods or conditions, leaving a gap in understanding the comprehensive effects of content balancing across diverse testing scenarios.

Our study addresses this gap by conducting a detailed simulation analysis of content balancing's impact on measurement precision under varying conditions, including different termination rules, sample sizes, and ability estimation methods. This study seeks to answer the following questions:

1. In computerized adaptive testing applications, when content balancing is not performed; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),

- Ability estimation methods (MLE, EAP)?

2. In computerized adaptive testing applications, when content areas are weighted equally for content balancing; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),
- Ability estimation methods (MLE, EAP)?

3. In computerized adaptive testing applications, when content areas are weighted disproportionately for content balancing; how do measurement precision and ability estimations change according to

- Termination rules (20 items, 60 items, SE≤.30),
- Sample sizes (N=250, N=500),
- Ability estimation methods (MLE, EAP)?

Methods

Research Model

This study aims to examine how content balancing in CAT applications with dichotomous items affects measurement precision under different conditions. The nature of this research is descriptive and simulative.

Data Generation

Participants for the CAT application were simulated using the R Studio program by the researcher (R Core Team, 2013). Initially, ability parameter values (true θ) for individuals were obtained, followed by item parameter values. Samples of two different sizes, 250 and 500 individuals, were created. The ability parameters of the individuals taking the test were generated to display a normal distribution θ~N(0, 1) within the range of -3 to +3 for each sample size condition.

The item pool for the CAT applications was created according to the 3PLM using the R Studio program. The item parameters were determined by the researcher to follow a uniform distribution. Feinberg and Rubright (2016) noted that item parameters are often simulated to follow a uniform distribution when using the three-parameter logistic model. For content balancing, item pools consisting of 750 items from five different content areas were created, weighted equally and disproportionately, using the 3PLM. In the item pool where content areas were weighted equally, each content area consisted of 150 items. In the item pool where content areas were weighted disproportionately, the different content areas contained 50, 50, 150, 250, and 250 items, respectively.

CAT applications yield better results when the items in the item pool have a sufficient number and a uniform distribution that caters to different ability levels and when the items are highly discriminative (DeMars, 2010; Flaugher, 2000). Therefore, item discrimination parameters “a” (ranging from 0.5 to 2), item difficulty parameters “b” (ranging from -3 to 3) and guessing parameters “c” (.05 to .2) were generated to follow a uniform distribution (Ree & Jensen, 1983; Thompson, 2009).

CAT Conditions

When no prior information about an individual's ability is available, assuming an average ability level is the most appropriate estimate. Starting the CAT application with an item of average difficulty level will be more psychometrically effective (Mills & Stocking, 1996). Therefore, the method within the range -.50<b<.50 was used as the test initiation rule for the simulative CAT application.

One of the best-known and oldest item selection methods, Maximum Fisher Information (MFI), involves selecting and administering an item that has the maximum Fisher information at a certain condition based on the test items previously administered (Han, 2018; Kalender, 2009). The MFI item selection method was chosen as a fixed condition in the simulation study. In the literature, there are various ability estimation methods based on dichotomous items and unidimensional IRT. The most frequently used among these methods are the MLE method and the Bayesian estimation method EAP (Chen et al., 1998; Segall, 2005). These two methods were considered as conditions for ability estimation in the current study. Fixed-length (20 and 60 items) and ability level's standard error (SE≤.30) rules were determined as conditions for test termination. To observe the performance of content balancing in short and long tests and to ensure adequate representation of all content areas, fixed test lengths of 20 and 60 items have been chosen. Among the methods proposed for content balancing while maintaining test efficiency, the most frequently used, simple, and understandable method is the CCAT method (Kingsbury & Zara, 1989). In the current simulation study, the CCAT method available in the "catR" package used for data analysis was employed as the content balancing method, leaving other content balancing methods outside the scope of this study. No item exposure rate control was conducted in the CAT application. The CAT conditions determined within the scope of the study are provided in Table 1.

Table 1
Conditions for the Computerized Adaptive Testing Application

CAT Components	Conditions	Number of Conditions
Termination Rule	20 items 60 items SE≤.30	3
Sample Size	250 500	2
Ability Estimation Method	MLE EAP	2
Test Initiation Rule	-.50<b<.50	1
Item Selection Method	MFI	1
Item Exposure Control	None	1
Content Balancing	None Equally Weighted Contents Differentially Weighted Contents	3

In the study, a total of 36 simulation conditions were examined, encompassing 3 termination rules, 2 sample sizes, 2 ability estimation methods, and 3 content balancing scenarios.

Data Analysis

In the scope of the research, measurement precision for each condition was evaluated using fidelity coefficient, RMSE (Root Mean Squared Error), and bias values. For most IRT studies, Harwell et al. (1996) recommended at least 25 replications to reduce sample bias and obtain stable and highly reliable results, but they also noted that in some studies this number may be much higher. These values were calculated separately for each of the 50 replications and then averaged.

The fidelity coefficient was assessed by calculating the correlation between the true θ levels, which were simulated at the start for individuals, and the θ levels estimated in each research condition and replication. The average correlation of the estimated θ values for each participant was obtained by averaging these correlations. The Pearson's correlation coefficient was used to calculate the fidelity coefficient, which is computed using the following formula:

$$r = \frac{\operatorname{cov}(\hat{\theta}, \theta)}{ss(\hat{\theta})ss(\theta)}$$

RMSE, the square root of the average of the squared differences between the estimated parameter value for each item in each replication and the true parameter value, is one of the most commonly used measures to evaluate the accuracy of estimates. It shows how far the estimates deviate from the true values using the Euclidean distance. Bias, indicating the systematic error related to the estimate, is equal to the difference between the average of the estimated parameter values for each item in each replication and the true parameter value, and is another measure indicating the precision of measurement. RMSE and bias values are calculated using the following equations (Zheng & Chang, 2014):

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

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_i\right)$$

In the equations, n represents the number of individuals, θi represents the individual's true ability level, and θi represents the estimated ability level of the individual. A high fidelity coefficient and low values of bias and RMSE indicate that there is no difference between the true ability level and the estimated ability level. The average test length in conditions where the termination criterion was set as SE≤.30 has also been examined.

To provide insights into test security, the maximum item exposure rates (r_{max}) for each condition were also examined.

Results

In this study, the RMSE, bias, and fit values calculated as indicators of measurement precision under 36 different conditions, along with the average test length in conditions where SE≤.30, are provided in Table 2.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(4); 395-407

Tablo 2
The Impact of Content Balancing on Measurement Precision Under Different Measurement Conditions in Computerized Individualised Testing Applications

Sample Size	Ability Estimation Method	Termination Rule	Content Balancing											
			None				Equally Weighted				Differentially Weighted			
			RMSE	*Bias*	*Correlation*	*Average Number of Items*	*RMSE*	*Bias*	*Correlation*	*Average Number of Items*	*RMSE*	*Bias*	*Correlation*	*Average Number of Items*
250	EAP	20 items	0.1900	0.0368	0.9830	-	0.1935	0.0322	0.9821	-	0.1954	0.0408	0.9795	
250	EAP	60 items	0.1282	0.0463	0.9931	-	0.1283	0.0462	0.9930	-	0.1270	0.0470	0.9922	
250	EAP	SE≤0.30	0.2025	0.0346	0.9805	17.40	0.2028	0.0299	0.9802	18.00	0.2020	0.0386	0.9778	17.97
250	MLE	20 items	0.2044	0.0357	0.9806	-	0.2135	0.0394	0.9791	-	0.2121	0.0486	0.9776	
250	MLE	60 items	0.1327	0.0453	0.9927	-	0.1349	0.0463	0.9925	-	0.1327	0.0494	0.9919	
250	MLE	SE≤0.30	0.2045	0.0348	0.9804	19.01	0.2070	0.0407	0.9801	19.73	0.2066	0.0453	0.9785	19.60
500	EAP	20 items	0.1868	0.0313	0.9821	-	0.1928	0.0325	0.9825	-	0.1962	0.0392	0.9797	
500	EAP	60 items	0.1255	0.0459	0.9929	-	0.1281	0.0453	0.9931	-	0.1287	0.0486	0.9923	
500	EAP	SE≤0.30	0.1968	0.0289	0.9799	17.32	0.2021	0.0314	0.9807	18.02	0.2011	0.0343	0.9783	18.14
500	MLE	20 items	0.2068	0.0385	0.9805	-	0.2132	0.0416	0.9796	-	0.2126	0.0503	0.9780	
500	MLE	60 items	0.1315	0.0465	0.9929	-	0.1329	0.0474	0.9928	-	0.1344	0.0537	0.9921	
500	MLE	SE≤0.30	0.2067	0.0374	0.9804	18.94	0.2097	0.0390	0.9800	19.67	0.2082	0.0464	0.9785	19.75

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications, *Journal of Measurement and Evaluation in Education and Psychology*, *15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

For different sample sizes, the estimated ability levels obtained by applying two different ability estimation methods were compared with the individuals' true ability levels in terms of RMSE and bias values.

When it comes to Table 2, bias values were close to zero in all conditions. The highest bias values (0.05) for both sample sizes were obtained when 60 items were used as the test termination rule in the EAP ability estimation method. When the test was terminated at 20 items and with the standard error termination rule, bias values (0.03) were found to be quite close to each other. Similarly, in the MLE ability estimation method, the highest bias value (0.05) for both sample sizes was obtained when 60 items were used as the test termination rule, both when content balancing was not performed and when content areas were equally weighted. When content balancing was performed with differentially weighted content areas, all bias values were relatively high (0.05) compared to other conditions. When the test was terminated at 20 items and with the standard error termination rule, bias values (0.04) were found to be quite close to each other. Generally, bias values were slightly higher in the MLE estimation method compared to the EAP method. Content balancing with equally weighted content areas did not affect bias values in both estimation methods when all conditions were considered together. Additionally, it was observed that bias values slightly increased in conditions of content balancing with differentially weighted content areas.

RMSE values were lowest when the test terminated at 60 items across all conditions, while they were similar for the standard error termination rule and when the test terminated at 20 items. Using the EAP ability estimation method, the lowest RMSE value (0.13) for both sample sizes was obtained when the test terminated at 60 items. When the test terminated at 20 items (0.19) and with the standard error termination rule (0.20), RMSE values were quite close to each other. Content balancing with equally and differentially weighted contents did not cause a significant change in RMSE values. Similarly, when using the MLE ability estimation method, the lowest RMSE value (0.13) was obtained when the test terminated at 60 items. When the test terminated at 20 items and with the standard error termination rule, RMSE values (0.21) were quite close to each other. Generally, RMSE values were slightly higher in the MLE estimation method compared to the EAP method. In conditions using the EAP ability estimation method, it was observed that RMSE values slightly decreased in larger samples when content balancing was not performed, while values were very close to each other when content balancing was performed. In conditions using the MLE ability estimation method, RMSE values were quite close to each other in small and large samples, whether content balancing was performed or not, and regardless of whether content areas were equally or differentially weighted (Table 2).

Correlations (r) between true and estimated ability levels were examined separately for two different sample sizes, three different termination rules, and content ratios used in content balancing, using different ability estimation methods. Accordingly, the highest correlation (r=0.99) between true and estimated ability levels for both sample sizes was obtained when the test terminated at 60 items, using both the EAP and MLE ability estimation methods. The fidelity coefficients obtained when the test was terminated at 20 items and with the standard error termination rule (SE≤.30) were quite close to each other. Content balancing did not affect the fidelity coefficients. It was observed that fidelity coefficients were slightly lower in conditions with differentially weighted contents compared to equally weighted content balancing (Table 2).

The effectiveness of whether content balancing was performed or not was also compared in terms of average number of items used in two different ability estimation methods. When the standard error termination rule (SE≤.30) was applied, the lowest average number of items (17.32) was obtained in conditions where the EAP ability estimation was used and content balancing was not performed. The average number of items was quite close across different sample sizes. The highest average number of

* Measurement and Evaluation Specialist, Ministry of National Education, Ankara, Türkiye, ilkayocal83@gmail.com, ORCID ID: 0009-0004-2246-6909
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara, Türkiye, nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Üçgül Öcal, İ., & Doğan, N. (2024). Effect of content balancing on measurement precision in computer adaptive testing applications, *Journal of Measurement and Evaluation in Education and Psychology, 15*(4), 394-407. https://doi.org/10.21031/epod.1438977

Received: 18.02.2024
Accepted: 18.12.2024

items was (19.75) in conditions where the MLE ability estimation was used and content balancing was performed with differentially weighted content areas. The average number of items was quite close across different sample sizes. Sample size did not affect the average number of items in either ability estimation method. Content balancing, in conditions with equally and differentially weighted content areas, increased the average number of items by approximately one item in conditions using both the EAP and MLE ability estimation methods.

The impact of content balancing on test security was also compared in terms of maximum item exposure (r_{max}) rates. The maximum item exposure rates obtained under 36 different conditions considered in the study are provided in Table 3.

Table 3

Maximum Item Exposure Rates (r_{max}) Under Different Conditions in Computerized Adaptive Testing Applications

Sample Size	Ability Estimation Method	Termination Rule	Content Balancing		
			None	Equally weighted	Differently weighted
250	EAP	20 items	0.5835	0.5593	0.6222
250	EAP	60 items	0.6767	0.6718	0.7006
250	EAP	SE≤0.30	0.5766	0.5428	0.5789
250	MLE	20 items	0.5580	0.5200	0.5521
250	MLE	60 items	0.6598	0.6526	0.6782
250	MLE	SE≤0.30	0.5522	0.5190	0.5382
500	EAP	20 items	0.5783	0.5349	0.6628
500	EAP	60 items	0.6704	0.6347	0.7311
500	EAP	SE≤0.30	0.5672	0.5225	0.6164
500	MLE	20 items	0.5370	0.5156	0.5727
500	MLE	60 items	0.6414	0.6326	0.6987
500	MLE	SE≤0.30	0.5334	0.5076	0.5696

In the small sample size, both the EAP and MLE ability estimation methods have shown that applying the termination at 20 items and the standard error termination rule (SE≤.30) reduced the maximum item exposure rate when content areas were equally weighted in content balancing. However, in the termination rule of stopping the test at 60 items, the rates are quite close to each other. In conditions where content balancing was done with differentially weighted content areas, the maximum item exposure rates increased with the EAP ability estimation method, whereas a decrease in this rate was observed when the MLE estimation method was used with the termination at 20 items and the standard error termination rule (SE≤.30).

In the large sample size, for both ability estimation methods, the maximum item exposure rates decreased in all conditions when content balancing was done with equally weighted content areas. In the case of content balancing with differentially weighted content areas, these rates increased in all conditions. Considering all conditions, the lowest item exposure rate (0.51) was observed in the large sample using the MLE estimation method with the standard error termination rule applied and when content areas were equally weighted in content balancing. The highest item exposure rate (0.73) was observed in the large sample using the EAP estimation method with the test termination rule at 60 items and when content balancing was done with differentially weighted content areas.

Discussion

Considering all findings obtained from the study, it has been observed that bias values, one of the indicators of measurement precision, slightly increase when content balancing involves differentially weighting content areas compared to other conditions. Generally, bias values were found to be lower in the EAP estimation method than in the MLE method. The RMSE value was not affected by whether content balancing was performed with equally or differentially weighted content areas when using the MLE estimation method. Without content balancing, RMSE values were quite close to each other in both small and large samples, regardless of whether content areas were equally or differentially weighted. Additionally, in conditions using the EAP estimation method, a slight decrease in RMSE values in larger samples was observed when no content balancing was performed. Generally, RMSE values were found to be slightly higher in the MLE method compared to the EAP method. An increase in the number of items reduced both RMSE and bias values, and the standard error termination rule and the termination at 20 items rules provided similar results. Regardless of sample size and ability estimation method, the highest correlation between true and estimated ability levels was obtained when the test terminated at 60 items. The selection of a 60-item test length in our study is supported by similar research and offers several advantages. Kingsbury et al. (2009) demonstrated that a 60-item exam allows for comprehensive content coverage and reliable, valid scores, equivalent to traditional tests of twice the length. Moreover, Sarı (2019) showed that longer tests mitigate adverse effects related to test security and reliability. Therefore, the 60-item length ensures adequate content representation and maintains high test reliability and validity, aligning with our study's goals. When content areas were differentially weighted, fidelity coefficients were found to be relatively lower compared to equal weighting. In both ability estimation methods, an increase in test length of about one item was observed when the standard error termination rule was applied. From this, it can be said that the increase in test length when content balancing is performed does not reduce test reliability to a significant extent. In all conditions, content balancing with equally weighted content areas reduced the maximum item exposure rates. Moreover, in the small sample, except for conditions where the test was terminated at 20 items and according to the SE<0.30 rule with the MLE method, content balancing with differentially weighted content areas increased the maximum item exposure rates. It can be said that content balancing conditions with equally weighted content areas perform better in terms of test security.

In synthesizing the outcomes of this study with those from related research, it's evident that the field of CAT is actively exploring the balance between measurement precision and content diversity. This study, alongside those by Leung et al.(2003b), Yasuda and Hull (2021), Yi and Chang (2010), and Zheng et al. (2013) collectively underscores the nuanced yet critical importance of content balancing in enhancing CAT's efficiency and accuracy without compromising item pool security and utilization. This study contributes to this body of knowledge by demonstrating that content balancing, while slightly increasing test length, does not detrimentally impact measurement precision. This finding aligns with Leung et al.'s (2003b) observation that certain item selection methods, notably the b-blocking method and MMM, optimize item pool utilization and minimize item overlap, suggesting that a thoughtful integration of stratification strategies and content balancing methods can achieve optimal outcomes in CAT applications. Moreover, the outcomes from Zheng et al. (2013) and Yasuda and Hull (2021) further reinforce the potential of content balancing strategies, such as the MMM, to effectively manage item exposure rates while maintaining test precision. This is particularly relevant in contexts requiring strict content specifications, where balancing can mitigate the risk of item overexposure without sacrificing measurement accuracy. Yi and Chang's (2010) introduction of a content-blocking method offers an innovative approach to item pool stratification, achieving balanced item usage and maintaining precision, which echoes this study's emphasis on the feasibility of content balancing in practical CAT designs. The collective findings suggest that while methodologies and focus areas may vary, the overarching goal remains consistent: refining CAT strategies to preserve the integrity of the testing process, optimize item pool usage, and ensure accurate and efficient measurement of abilities.

Comparing the outcomes of various studies on CAT, we observe diverse approaches and impacts of content balancing on measurement precision. Leung et al. (2003b) highlight how specific item selection methods like b-blocking method to multiple stratification and MMM optimize item pool utilization without affecting measurement accuracy, contrasting with our study's emphasis on the slight increase in test length due to content balancing. Zheng et al. (2013) and Yasuda and Hull (2021) focus on content balancing's effect on specific domains or inventories, showcasing its variable impact on measurement precision. Yi and Chang's (2010) content-blocking method presents a novel approach, differing from traditional strategies by enhancing item pool usage efficiently. These differences underline the complexity of optimizing CAT, suggesting that the choice of content balancing strategy should be tailored to specific testing requirements and goals.

This study examined the effect of content balancing on measurement precision in dichotomous items under different measurement conditions in CAT applications. Control of item exposure rate, which holds significant importance in the CAT algorithm, was beyond the scope of this study. Future research could examine the impact of content balancing on measurement precision with control of item exposure rate. Similarly, the effect of content balancing when using different item selection methods could be explored. In the current study, the CCAT method was used as the content balancing method. Future studies could compare the performance of other content balancing methods on measurement precision using different packages or software.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Author Contribution: İlkay ÜÇGÜL ÖCAL: conceptualization, investigation, methodology, data simulation, data analysis, supervision, writing - review & editing. Nuri DOĞAN: conceptualization, methodology, writing - original draft, formal analysis.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as data has been simulated in this study.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Chen, S. K., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, *58*(4), 569.

Chen, S.Y., & Ankenmann, R. D. (2004). Effects of Practical Constraints on Item Selection Rules at the Early Stages of Computerized Adaptive Testing. Journal of Educational Measurement, *41*(2), 149–174. http://www.jstor.org/stable/1435211

Cheng, Y., & Chang, H-H. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*(6), 467-482.

DeMars, C. (2010). *Item response theory*. Oxford Academic. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Demir, S. (2019). *Bireyselleştirilmiş bilgisayarlı sınıflama testlerinde sınıflama doğruluğunun incelenmesi*. [Doctoral Dissertation, Hacettepe Üniversitesi].

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

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36-49. https://doi.org/10.1111/emip.12111

Flaugher, R. (2000). Item pools. In H. Wainer (Ed.), *Computerized Adaptive Testing: A primer* (2nd ed., pp. 37-60). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications*. Boston: Kluwer.

Han, K.T. (2018). Components of the item selection algorithm in computerized adaptive testing. *Journal of Educational Evaluation for Health Professions*, *15*(7). https://doi.org/ 10.3352/jeehp.2018.15.7

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

Kalender, İ. (2009). Başarı ve yetenek kestirimlerinde yeni bir yaklaşım: Bilgisayar ortamında bireyselleştirilmiş testler (Computerized adaptive tests-CAT). *CITO Egitim Kuram ve Uygulama*, *5*, 39-48.

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

Kingsbury, G. G., Bontempo, B., & Zara A. R. (2009). *A comparison of CAT with LOFT methods for certification examinations.* [Conference presentation]. NOCA Annual Educational Conference.

Leung, C. K., Chang, H. H., & Hau, K. T. (2000). *Content balancing in stratified computerized adaptive testing designs* [Paper presentation]. AERA Annual Meeting, New Orleans.

Leung, C.K., Chang, H.H., & Hau, K.T. (2003a). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment*, *2*(5).

Leung, C.K., Chang, H.H., & Hau, K.T. (2003b). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, *63*(2), 257-270.

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

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*(4), 287–304. https://doi.org/10.1207/s15324818ame0904_1

Özdemir, B., & Gelbal, S. (2015). İçerik ağırlıklandırmasının maddeler-arası boyutluluk modeline dayalı çok boyutlu bilgisayar ortamında bireyselleştirilmiş test yöntemleri üzerindeki etkisinin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2). https://doi.org/10.21031/epod.03278

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

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters, In Weiss, D. (Ed). *New horizons in testing latent trait test theory and computerized adaptive testing*, 135-146. London: Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50017-2

Rudner, L. M. (1998). An online, interactive, computer adaptive testing tutorial. Retrieved December 25, 2023, from https://edres.org/scripts/cat/catdemo.htm

Sari, H. İ., & Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5). https://doi.org/10.12738/estp.2017.5.0484

Sarı. H. İ. (2019). Investigating consequences of using item pre-knowledge in computerized multistage testing. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi,* 39(*2*), 1113-1134. https://doi.org/10.17152/gefad.535376

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*. New York: Academic Press.

Song, T. (2010). *The effect of fittıng a tridimensional irt model to multidimensional data in content-balanced computerized adaptive testing*. [Doctoral Dissertation, Michigan State University].

Şahin, A., Özbaşı D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive tests. *Eurasian Journal of Educational Research*, *17*(69), 21-36. http://dx.doi.org/10.14689/ejer.2017.69.2

Şenel, S. (2021). *Bilgisayar ortamında bireye uyarlanmış testler*. Pegem Akademi, Ankara.

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., & Weiss, D. A. (2019). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/wqzt-9427

Tian, J., Miao, D., Zhu, X., & Gong, J. (2007). An introduction to the computerized adaptive testing. *Us-China Education Review*, *4*(1), 72-81.

Van der Linden, W., & Glas, G. A. W. (2002). *Computerized adaptive testing: theory and practice*. Kluwer Academic Publishers.

Yasuda, J. I., & Hull, M. M. (2021). *Balancing content of computerized adaptive testing for the Force Concept Inventory* [Conference presentation]. Physics Education Research.

Yi, Q., & Chang, H-H. (2010). a-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359-378. https://doi.org/10.1348/000711003770480084

Zheng, Y., Chang, C-H., & Chang, H-H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Qual Life Res, 22*(3), 491-499. https://doi.org/ 10.1007/s11136-012-0179-6

Zheng, Y., & Chang, H-H. (2014). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*. *39*(2), 105-118. https://doi.org/10.1177/0146621614544519