Measuring language ability of students with compensatory multidimensional CAT: A post-hoc simulation study

Burhanettin Ozdemir¹ · Selahattin Gelbal²

Received: 26 April 2021 / Accepted: 5 December 2021 / Published online: 5 January 2022
© The Author(s) 2022

Abstract

The computerized adaptive tests (CAT) apply an adaptive process in which the items are tailored to individuals' ability scores. The multidimensional CAT (MCAT) designs differ in terms of different item selection, ability estimation, and termination methods being used. This study aims at investigating the performance of the MCAT designs used to measure the language ability of students and to compare the results of MCAT designs with the outcomes of corresponding paper–pencil tests. For this purpose, items in the English Proficiency Tests (EPT) were used to create a multi-dimensional item pool that consists of 599 items. The performance of the MCAT designs was evaluated and compared based on the reliability coefficients, root means square error (RMSE), test-length, and root means squared difference (RMSD) statistics, respectively. Therefore, 36 different conditions were investigated in total. The results of the post-hoc simulation designs indicate that the MCAT designs with the A-optimality item selection method outperformed MCAT designs with other item selection methods by decreasing the test length and RMSD values without any sacrifice in test reliability. Additionally, the best error variance stopping rule for each MCAT algorithm with A-optimality item selection could be considered as 0.25 with 27.9 average test length and 30 items for the fixed test-length stopping rule for the Bayesian MAP method. Overall, MCAT designs tend to decrease the test length by 60 to 65 percent and provide ability estimations with higher precision compared to the traditional paper–pencil tests with 65 to 75 items. Therefore, it is suggested to use the A-optimality method for item selection and the Bayesian MAP method for ability estimation for the MCAT designs since the MCAT algorithm with these specifications shows better performance than others.

Keywords Computerized adaptive testing · Language testing · Post-hoc simulation · Multidimensional IRT

✉ Burhanettin Ozdemir
bozdemir@psu.edu.sa

Extended author information available on the last page of the article

1 Introduction

Language testing is a broad area that requires sophisticated methods to measure language skills or abilities such as reading, listening, and speaking skills. Traditional measurement methods used to measure individuals' language skills, seem to be insufficient which has led to the development of more complex and advanced testing methods. Therefore, there has been a huge development in this area of study.

Traditional fixed test length paper–pencil tests are used, in general, to measure traits or abilities with a restricted range. Moreover, most of the items in these tests are more suitable for examinees with average ability levels. Another shortcoming of a traditional paper–pencil test is that there is a time limit for responding to all items which causes inaccurate ability estimations for those who respond to items more slowly than other examinees (Weiss, 2005). On the other hand, Computerized adaptive tests (CATs) match the properties of items with each examinee's ability level to obtain more accurate estimates.

Due to the adaptive nature of the measurement process of CAT designs, the very easy and difficult items are eliminated for each test-taker which decreases the test length and performance times (Curi & Silvia, 2019; Sukamolson, 2002). Thus, CATs are assumed to be advantageous compared to traditional paper–pencil exams. Because they can decrease the number of items while increasing or maintaining the same level of measurement quality as the corresponding paper–pencil tests. Therefore, considering these advantages, it is obvious that the CATs make the testing process more effective and efficient.

One of the most important advantages of CAT is that it provides more reliable measures with a shorter test (Wainer, 1993). The main reason behind this is that the items that provide the highest information at the currently estimated ability are selected due to the adaptive nature of the process. Besides, CATs are more flexible to testing time and have the feature of providing results of the test as soon as the testing process is terminated (Curi & Silvia, 2019; Lin, 2012; Weiss, 1983). Moreover, CATs based on item response theory (IRT) provide comparable ability scores between the test takers answering a different set of items and taking the test at different times (Curi & Silvia, 2019; Kreitzberg et al., 1978; Wainer, 2000).

Developments in computer technologies, along with item response theory (IRT) models increased the applicability of unidimensional and multidimensional computerized adaptive tests (Lee et al., 2019; Wang & Chen, 2004). Adaptive tests using unidimensional IRT-based item selection and theta estimation methods are called Unidimensional CAT (U-CAT) methods, while adaptive tests using multidimensional IRT-based item selection and theta estimation methods are called Multidimensional CAT (MCAT) methods (Wang & Chen, 2004). Although CAT designs mostly benefit from unidimensional IRT methods, they might not be applicable in real test situations.

Especially, measuring cognitive abilities, reading and writing skills, performance tasks, and clinical abilities require using multi-dimensional IRT (MIRT)

models (van der Linden & Hambleton, 1997, pp. 221). Additionally, the MCAT applications of a multidimensional test have been considered to be more efficient and advantageous compared to the unidimensional CAT applications (Chien & Wang, 2017; Lee et al., 2019; Wang, 2010). Therefore, increasing trends in using MIRT models, and considering the adaptive test as a more reliable alternative to traditional paper–pencil tests lead us to the development of MCAT procedure that combines both MIRT models and CAT procedures (Segall, 1996, 2001).

The post-hoc CAT simulation methods, on the other hand, allow comparing the performance and the outcomes of the CAT designs to the corresponding paper–pencil versions (e.g., Wang et al., 1999; Weiss, 2005). The post-hoc simulations provide preliminary analyses that allow investigating the adaptive test performance of a given test administered in paper–pencil format. Thus, post-hoc CAT simulations allow examining how much reduction is achieved and to what extent the standard error associated with ability parameters is decreased (Kalender, 2011).

1.1 The Computerized Adaptive Testing (CAT) Process

The CAT systems, in general, consist of five important blocks (Weiss & Kingsbury, 1984): item bank, starting rule, item selection, scoring method, and termination (or stopping) rule (Luo et al., 2020). The first step in a CAT process is to construct an item bank that contains the previously administered items calibrated with measurement models. The item parameters, such as difficulty, discrimination and guessing parameters, are obtained from the item calibration process.

The CAT process workflow is presented in Fig. 1 (Oppl et al., 2017, p. 4, Fig. 1). The process is initiated by selecting the first item from the item bank based on the predefined criteria. After administering the first item, the ability score is estimated with an adopted scoring method based on IRT models, and a new item is selected from the item bank in accordance with the individual's previously estimated ability level. Typically, a more challenging item, that fits best to the estimated ability score, is selected if the examinee answered the previous items correctly, and vice versa. The item selection and ability estimation cycle (step 2 to step 5 in Figure-1) is repeated until the predefined termination rule is met. The CAT process is usually terminated based on fixed test-length or precision-based stopping rule (Segall, 2004).

Fig. 1 The computerized adaptive testing (CAT) procedure

Post-hoc CAT simulations give insight into which CAT designs yield more consistent results compared to the corresponding paper–pencil test. Additionally, it provides a more accurate psychometric characteristic of given examinees (Wang et al., 1999) since real data sets and item parameters are used when testing the performance of MCAT designs. More detailed information about the components of MCAT designs used in this study, such as IRT models, ability estimation and item selection methods, content balancing, and test termination rules are provided in the following sections.

1.2 Multidimensional IRT models

Multidimensional item response theory (MIRT) models are generally classified as Compensatory and Non-compensatory models (Sijtsma & Junker, 2006). An individual's relatively low score from one dimension can be compensated by high scores from other dimensions in compensatory MIRT models. However, this compensation is not possible for non-compensatory models. Therefore, which method to use depends on the structure of ability or skill being measured by the test. If the relation between ability parameters obtained from different dimensions is high or ability parameters can be compensated by other higher ability parameters, then compensatory MIRT models should be utilized.

In this study, a two-parameter logistic MIRT (2PL MIRT) model, in which pseudo chance parameters are set equal to zero (0) was used, since pseudo chance parameters of items included in the item bank ranged between 0 and 0.10. Therefore, the 2PL MIRT model was favored to decrease calculation time. The formula for this model is as follows:

$$P(U_{ij} = 1 | \theta_j, \mathbf{a}_i, d_i) = \frac{e^{(\mathbf{a}_i \theta_j' + d_i)}}{1 + e^{(\mathbf{a}_i \theta_j' + d_i)}} \quad (1)$$

where θ_j represents ability parameters of m dimensions and is a $1 \times m$ vector. Likewise, $\mathbf{a}_i$ is a $1 \times m$ discrimination vector and b_i and d_i are scalars, representing item difficulty and item easiness, respectively.

1.3 Item and Test information functions

Information functions occupy an important place in the item response theory model. They provide information about the accuracy of estimated ability parameters and the amount of error during the process of measurement. Since the items in a test provide information about the ability being measured to a certain degree, the amount of information provided by the items depends on how consistent the estimated abilities and item parameters are. Therefore, item information functions are used to estimate the information provided by each item at different estimated theta values.

The mathematical formula for estimating item information was first proposed by R.A. Fisher in 1925 (Kullback, 1959). The Fisher's information function formula is given below:

$$I(\theta) = E_{\theta} \left[\frac{d \log f(Y; \theta)}{d\theta} \right]^2 \quad (2)$$

where θ denotes ability parameters and $f(Y; \theta)$ denotes probability density function of observed score Y . Fisher has shown that the asymptotic variance of θ estimated with MLE (maximum likelihood estimation) is equal to the multiplicative inverse of $I(\theta)$. It also provides a lower bound of the variance of θ estimated by other unbiased estimation methods.

Fisher's test information function is used to calculate the total test information for all items administered at each θ value. The following formula is Fisher's test information function:

$$I_j(\theta_j) = \sum_j E_{\theta} \left[\frac{d \log f(Y; \theta)}{d\theta} \right]^2 \quad (3)$$

The most important feature of the test information function is that it provides information about the degree of accuracy of estimated θ and enables us to calculate measurement errors.

1.4 Item selection methods

The first studies investigating item selection methods for multidimensional adaptive tests were conducted by Bloxom and Vale (1987). They adapted the item selection method based on Bayesian approximation procedures proposed by Owen (1969, 1975) for unidimensional CAT to multidimensional CAT procedures. The study of Bloxom and Vale (1987) about MCAT was followed by other researchers, such as van der Linden (1996, 1999), Segall (1996, 2000), Fan and Hsu (1996), and Veldkamp and van der Linden (2002).

Some of the item selection methods developed for multidimensional CAT utilize the optimal designs. Optimal design methods are used to optimize statistical inferences by calculating determinants or the covariance-matrix trace and information matrix. The most commonly used optimal designs are D-optimality, A-optimality, C-optimality, and E-optimality (Silvey, 1980, p. 10).

A-optimality minimizes the composite measure's error variance (van der Linden, 1999), while C-optimality and Kullback–Leibler information (KLI) methods (Veldkamp & van der Linden, 2002) maximize the prior distribution. Moreover, D-optimality is obtained from the determinant of the posterior information matrix (Luecht, 1996; Segall, 1996). Besides, these optimal designs are commonly used in different areas such as educational and medical sciences (Berger & Wong, 2005). Fisher's information matrix is of great importance in these designs since the Fisher's information matrix is used to calculate information about latent variables that explain the observed variables.

A-optimality and D-optimality item selection methods differ in terms of calculating the determinant of the information matrix. Moreover, A-optimality takes the variance of other ability parameters into account, when selecting the next item. The

following section provides information about the ability estimation parameters used in this study.

1.5 Ability estimation methods

1.5.1 MLE Ability Estimation Method

Maximum likelihood estimation (MLE) aims to calculate the most likely ability score of the examinees given the responses to items in a test. When estimated item parameters are given, $L(U_j|\theta_j)$ denotes the likelihood function of an examinee's response pattern. The formula for likelihood function is as follows:

$$L(U_j|\theta_j) = \prod_{i=1}^n P(U_{ij}|\theta_j) \quad (4)$$

where u_j represents responses of examinee j to items and θ_j denotes ability parameter of examinee j . Maximum likelihood estimation of ability parameter (θ_j) equals to the θ value that maximizes this likelihood function. The log-likelihood function's first derivative is calculated and solved for θ to find out the θ value that maximizes the likelihood function.

MLE, in general, has a larger standard error and RMSE values than Bayesian scoring methods (Wang & Vispoel, 1998; Warm, 1989; Weiss & McBride, 1984). Another disadvantage of MLE ability estimation is that it cannot estimate θ values when the examinees answer all items correctly or incorrectly at the beginning of the CAT procedure. The θ values can either be restricted with $[-4, 4]$ interval or Bayesian estimation methods are used (Song, 2010).

1.5.2 Fisher's Scoring Method

Another method commonly used to estimate examinees' ability parameters is called Fisher's scoring method. Compared to the MLE estimation method, this method utilizes the expected Fisher's information function instead of Fisher's exact information function. The expected Fisher's information function is denoted by $J(\theta)$ and the following formula is used to calculate it:

$$J(\theta; u_{ij}) = E[I(\theta; u_{ij})] \quad (5)$$

Therefore, the formula for Fisher's scoring method is as follows:

$$\theta_{k+1} = \theta_k + \frac{U(\theta; u_{ij})}{J(\theta; u_{ij})} \quad (6)$$

Although Fisher's scoring method requires more iterations compared to Newton–Raphson iteration methods, calculating the expected value of Fisher's information is easier than calculating Fisher's information. Therefore, Fisher's scoring method calculates ability parameters in a shorter time. Besides, the programming of Fisher's scoring method is easier compared to MLE estimation methods. A similar

formula is used when tests consist of more than one dimension. The only difference is that θ is a vector rather than a scalar when the test is multidimensional.

1.5.3 Bayesian Maximum A Posteriori (MAP)

Bayesian ability estimation methods utilize prior information about examinees' ability parameters. This virtual distribution is called the prior distribution of θ which has a normal distribution with the mean equals 0 and standard deviation 1 in the context of adaptive testing.

MAP (maximum a posteriori), EAP (expected a posteriori) and Owen's normal approximation methods are the most commonly used Bayesian estimation methods. The Bayesian EAP estimation method calculates the expected value of the posterior distribution. The expected value of the posterior distribution of ability parameter is calculated using the following formula:

$$\hat{\theta} = E(\theta|U_j) = \int_{-\infty}^{\infty} \theta h(\theta|U_j) d\theta \quad (7)$$

MAP estimation method, developed by Samejima (1969), takes the mode of the posterior distribution of ability parameters. In this study, MLE-based Fisher's scoring and Bayesian MAP methods were used to estimate examinees' ability parameters.

1.6 MCAT stopping rules

The Adaptive testing process is iterative and is terminated when the specified termination rule or condition is met (Reckase, 2009; Wainer, 2000). The computerized adaptive testing process is terminated when the predetermined reliability or precision level is achieved, or a fixed number of items is administered; or the ability parameters are estimated with a certain confidence interval (Yao, 2012).

Test length, or the number of items administered to each examinee, might vary in a large range when the precision-based stopping rule is utilized. Moreover, testing time can be longer compared to paper–pencil tests. When the fixed test length stopping rule is favored, then the applicability of adaptive testing might increase and the testing item might vary in an acceptable range. However, the desired precision level of estimated ability parameters may not be achieved for every examinee in this case. Thus, both fixed test length and precision stopping rule can be applied together to eliminate these shortcomings. Another stopping rule applied in the context of MCAT is based on error variance, in which tests are terminated when the error variances of estimated ability parameters decrease to a certain predetermined level.

In this study, the fixed test length and error variance stopping rules developed for MCAT procedures were used to determine the best stopping rule along with other conditions. For fixed test length, the three different test length conditions (30, 40, and 50) were tested, while for the error variance stopping rule, three different compensatory error variance stopping rule conditions (0.20, 0.25, and 0.30) were tested.

1.7 Content Balancing

The blueprints are developed to set a guideline for the test developer in the context of a traditional paper–pencil test considering the properties of the domains. On the other hand, since the most informative items are selected during the MCAT procedure, the distribution of selected items for different content may differ compared to its paper–pencil test counterpart. Therefore, examinees will answer a differing number of items related to each content. Therefore, the idea of content balancing was first proposed by Green and his colleagues (1984) to ensure the content validity and the balanced content distributions for each examinee in the context of adaptive testing.

Wainer and Kiely (1987) developed the testlet method instead of asking for independent items; Kingsbury and Zara (1989) developed the constrained CAT method which takes the distribution of items for contents of a paper–pencil test into consideration while selecting the most informative items. Moreover, Leung et al. (2000) developed the constrained content balancing for CAT which is based on the method developed by Kingsbury and Zara (1991); and Chen and Ankenmann (2004) developed the modified multinomial model (MMM) which enables content balancing (Song, 2010).

1.8 Purpose of study

The English Proficiency Test (EPT) is developed by Hacettepe University and administered in a paper–pencil (P&P) format. It is used to determine the language proficiency level of each freshman. Therefore, all the new students enrolled in the university must take it. Those who fail the EPT have to take English preparation classes until they can successfully pass it. The ultimate goal of this study is (a) to compare the performance of the different Multidimensional CAT designs and (b) to determine the most suitable MCAT design that suits the EPT, and (c) to compare the paper–pencil (P&P) test results to those of the new MCAT designs.

2 Research questions

Research questions of this study are as follows:

- (1) How do using different combinations of item selection, ability estimation methods, and stopping rules affect the MCAT performance indicators, such as RMSD, RMSE statistics, test length, and reliability indices?
- (2) Which item selection method provide more consistent and reliable results in the context of MCAT?
- (3) Which ability estimation method provides more consistent and reliable results given the same conditions in the context of MCAT?

- (4) What is the most suitable stopping rule for the MCAT version of EPT?
- (5) What are the advantages of developing the MCAT version of EPT compared to the original paper–pencil format according to post-hoc simulation study?

3 Methodology

Firstly, a multidimensional item bank (pool) was constructed using the items in various EPT tests administered between 2009 and 2013 that consist of three main sections that are reading, grammar, and listening. The minimum number of candidates that took these forms was 800 indicating the existence of an adequate response rate for each item. To construct the item pool, 628 items in total were calibrated with a multidimensional compensatory model with two parameters (MC-2PL) that allows items to load on more than one dimension (free-floating calibration).

Secondly, based on the item calibration results, the misfitting items that have low discrimination, items that have difficulty beyond the [-4, 4] interval, and items that have high guessing parameters were excluded. Finally, after excluding the poor items, a three-dimensional item pool consisting of 559 items in total was constructed. The final item pool contains 250 grammar items, 199 reading items, and 110 listening items, respectively. For each iteration, the standard normal distribution was used to generate theta parameters (θ) and the sample size was fixed to 500.

The MCAT designs differed in terms of item selection and ability estimation methods, and termination rules, respectively. Three different item selection methods that are D-Optimality, A-Optimality, and random items selection methods were employed. For ability estimation, MLE-based Fisher scoring and Bayesian-MAP methods were used to calculate test takers’ ability scores. As a stopping rule, the precision-based and fixed test-length termination rules were utilized.

The content-balancing was imposed by specifying target content distribution (0.45, 0.20, and 0.35) proportional to the number of items related to grammar, listening, and reading contents of EPT to ensure content validity. Along with all these conditions, the randomesque item exposure control technique was used to select one item among the 10 most informative items randomly at the current ability estimate to control the exposure rate for the informative items. All the post-hoc simulation conditions are listed in Table 1.

Table 1 presents the post-hoc simulation conditions including ability estimation methods, item selection methods, and stopping rules. In total, the performance of

Table 1 List of post-hoc simulation conditions

Ability (θ) estimation methods	Item selection methods	Stopping rules
<ul style="list-style-type: none">•Maximum Likelihood Estimation (MLE)•Fisher’s scoring method	<ul style="list-style-type: none">•A-optimality•D-optimality•Random (non- adaptive)	<ul style="list-style-type: none">•Fixed test length (30, 40, 50)•Standard Error stopping rule (0.20, 0.25, 0.30)

36 different conditions was tested and compared based on RMSD statistics between true theta and the theta obtained from each MCAT design, RMSE, reliability, and test-length (total number of items administered) for each MCAT design.

4 Findings

In this study, the performance of different MCAT conditions was compared with respect to test length, reliability, RMSE, and RMSD statistics, and findings were provided for each ability estimation method based on a different combination of item selection methods and stopping rules.

Table 2 presents the result of the MCAT with A-optimality item selection method and fixed test length stopping rule for each ability estimation method with balancing.

The results of MCAT with A-optimality item selection and fixed test length stopping rule in Table 2 indicate that the reliability coefficients of MCAT algorithm using Bayesian MAP and MLE based Fisher's estimation methods were similar and the difference was negligibly small. However, MCAT with the Bayesian MAP method had relatively smaller RMSE and RMSD statistics compared to Fisher's scoring method. Besides, as the number of items increased, reliability coefficients remained the same and RMSE and RMSD statistics tended to decrease somewhat. When the test length was increased from 30 to 40, both RMSD and RMSE statistics decreased somewhat. Therefore, the test length can be set as 30 for each MCAT algorithm with A-optimality item selection and fixed test length stopping rule.

The results of MCAT with A-optimality item selection and error variance stopping rule shown in Table 2 indicate that the MCAT designs using Bayesian MAP

Table 2 Result of MCAT with A-optimality item selection and different stopping rules (fixed vs standard error)

Scoring Methods	Stopping rule	Test length	Reliability			RMSE			RMSD		
			dim1	dim2	dim3	dim1	dim2	dim3	dim1	dim2	dim3
Fisher	Fixed										
	30	30	0.93	0.83	0.81	0.251	0.383	0.400	0.217	0.331	0.319
	40	40	0.95	0.86	0.84	0.230	0.360	0.377	0.194	0.293	0.284
	50	50	0.95	0.87	0.85	0.215	0.340	0.358	0.175	0.277	0.277
Bayesian	30	30	0.94	0.84	0.83	0.239	0.372	0.387	0.209	0.323	0.308
	40	40	0.95	0.86	0.84	0.219	0.350	0.368	0.178	0.294	0.284
	50	50	0.96	0.88	0.86	0.207	0.333	0.353	0.170	0.260	0.261
	SE		dim1	dim2	dim3	dim1	dim2	dim3	dim1	dim2	dim3
Fisher	0.20	49.7	0.99	0.98	0.98	0.219	0.342	0.359	2.495	2.417	2.340
	0.25	34.6	0.95	0.85	0.84	0.239	0.371	0.388	0.578	0.592	0.580
	0.30	22.5	0.94	0.84	0.82	0.278	0.415	0.428	0.329	0.376	0.368
Bayesian	0.20	45.5	0.95	0.87	0.85	0.210	0.339	0.358	0.174	0.267	0.275
	0.25	27.9	0.94	0.84	0.82	0.240	0.379	0.393	0.208	0.322	0.302
	0.30	17.7	0.91	0.79	0.78	0.284	0.416	0.427	0.266	0.379	0.361

The bold entries represent the results of optimal MCAT designs for given conditions

and MLE-based Fisher's ability estimation methods had similar results concerning reliability coefficients and RMSE statistics, respectively. However, MCAT with the Bayesian MAP method yielded smaller RMSD statistics compared to Fisher's scoring method. Moreover, reliability coefficients remained the same, and RMSE and RMSD statistics tended to decrease somewhat as the number of items increased. Therefore, the best error variance stopping rule for each MCAT algorithm with A-optimality item selection can be considered as 0.25 with 34.6 average test length for Fisher's scoring and 27.9 average test length for the Bayesian MAP method, respectively.

Figure 2 depicts the changes in RMSD, RMSE, and reliability statistics as a function stopping rules associated with the MCAT designs with A-optimality item selection for each ability estimation method.

Table 3 presents the result of MCAT algorithms with D-optimality item selection method for different combinations of ability estimations methods and stopping rules.

The result of MCAT with D-optimality item selection and fixed test length stopping indicate that the MCAT algorithm with the Bayesian MAP method yielded higher reliability coefficients, smaller RMSE, and RMSD statistics compared to Fisher's scoring method for each condition. The stopping rule can be set as 30 for the MCAT algorithm with D-optimality item selection and Bayesian MAP method; and 40 for MCAT algorithm with D-optimality item selection and Fisher's scoring method. Thus, the Bayesian MAP ability estimation method outperformed Fisher's ability estimation methods when D-optimality item selection methods and fixed test length stopping rules were used.

The result of MCAT with D-optimality item selection and error variance stopping rule shown in Table 3 indicates that Bayesian MAP ability estimation method outperformed MLE based Fisher's scoring method in terms of reliability, RMSE, RMSD statistics, and the average test-length for each error variance based stopping criterion. Therefore, the 0.25 error variance stopping rule can be considered as the best error variance stopping rule for Bayesian MAP ability estimation method with D-optimality item selection which resulted in 28 average test length, while 0.30 for Fisher's scoring method with 39.9 average test length, respectively.

Figure 3 depicts the change in RMSD, RMSE, and reliability statistics associated with the MCAT designs with D-optimality item selection for different ability estimation methods and stopping rules. The first column depicts the change in the correspondence statistics related to the MCAT algorithm with fixed test length stopping rules, while the second column represents the change in the correspondence statistics related to the MCAT algorithm with precision-based (standard error) stopping rules.

Table 4 presents the result of MCAT with D-optimality item selection, different stopping rules and ability estimation methods.

Table 4 presents the result of the MCAT with random item selection (non-adaptive item selection method) for different combinations of stopping rule ability estimation methods. The reason behind using random (or non-adaptive) item selection methods is to investigate if other item selection methods (A-optimality and D-optimality) caused any decrease in RMSE statistics and test-lengths independent of

Fig. 2 A-optimality MCAT results with different stopping rules

ability estimation methods, or if changes in these statistics were only due to ability estimation methods.

The result of MCAT with random item selection (non-adaptive item selection method) and fixed test length stopping rule shown in Table 4 indicates that reliability coefficients were substantially small, while RMSE and RMSD statistics were substantially large regardless of ability estimation methods being used. This indicates that the random item selection method was outperformed by the other two item selection methods. Additionally, the results of MCAT with random item selection methods were not as reliable and consistent as others. Therefore, one can conclude

Table 3 Result of MCAT with D-optimality item selection and different stopping rules (fixed vs standard error)

Scoring Methods	Stopping rule	Test length	Reliability			RMSE			RMSD		
	Fixed		dim1	dim2	dim3	dim1	dim2	dim3	dim1	dim2	dim3
Fisher	30	30	0.82	0.75	0.72	0.383	0.439	0.459	0.343	0.397	0.403
	40	40	0.86	0.79	0.77	0.346	0.407	0.428	0.320	0.372	0.387
	50	50	0.89	0.84	0.81	0.321	0.385	0.407	0.313	0.332	0.346
Bayesian	30	30	0.94	0.84	0.82	0.239	0.373	0.388	0.203	0.322	0.306
	40	40	0.95	0.86	0.84	0.219	0.350	0.367	0.185	0.292	0.286
	50	50	0.96	0.88	0.86	0.205	0.334	0.353	0.172	0.266	0.266
	SE		dim1	dim2	dim3	dim1	dim2	dim3	dim1	dim2	dim3
Fisher	0,20	50.0	0.89	0.83	0.80	0.319	0.385	0.407	0.291	0.326	0.329
	0,25	49.2	0.89	0.83	0.81	0.320	0.387	0.408	0.304	0.345	0.351
	0,30	39.9	0.87	0.80	0.78	0.334	0.405	0.424	0.321	0.364	0.363
Bayesian	0,20	45.8	0.96	0.87	0.86	0.210	0.340	0.357	0.180	0.284	0.278
	0,25	28.0	0.94	0.84	0.83	0.240	0.379	0.392	0.198	0.331	0.303
	0,30	17.7	0.92	0.80	0.78	0.283	0.417	0.427	0.272	0.381	0.361

The bold entries represent the results of optimal MCAT designs for given conditions

that A-optimality and D-optimality item selection methods bring about a significant increment in reliability and a significant decrement in RMSE and RMSD correspondence statistics.

The results of MCAT with random item selection and error variance test length stopping rule shown in Table 4 indicate that reliability coefficients were substantially small, while RMSE and RMSD statistics were substantially large regardless of ability estimation methods being used. This indicates that the random item selection method was outperformed by the other two item selection methods when the error variance stopping rule was used. Besides, increasing the error variance stopping rule from 0.20 to 0.30 did not cause a substantial decrease in the average number of items administered. Thus, using optimality-based item selection methods was more advantageous compared to the non-adaptive item selection method in terms of test length, RMSE, RMSD, and reliability coefficients.

Figure 4 depicts the RMSD, RMSE, and reliability statistics as a function of the fixed test length stopping rule associated with the MCAT designs with non-adaptive random item selection for each ability estimation method.

5 Conclusion and Discussion

In this study, the performance of the multidimensional computerized adaptive test (MCAT) designs using the different combinations of item selection, ability estimation methods, and termination rules were examined to find out the most effective MCAT design that could be used to measure individuals' language skills as an

Fig. 3 D-optimality MCAT results with different stopping rules

alternative to the paper–pencil version of EPT. To this end, a multidimensional item bank was constructed with the items administered in the various EPT forms. Moreover, the results of MCAT with and without content balancing were compared to the

Table 4 Result of MCAT with random item selection and different stopping rules (fixed vs standard error)

Scoring Methods	Stopping rule	Test length	Reliability			RMSE			RMSD		
			dim1	dim2	dim3	dim1	dim2	dim3	dim1	dim2	dim3
Fisher	Fixed										
	30	30	0.76	0.67	0.65	0.435	0.492	0.508	0.421	0.440	0.448
	40	40	0.95	0.93	0.92	0.389	0.453	0.472	1.503	1.433	1.434
Bayesian	50	50	0.85	0.76	0.74	0.354	0.425	0.444	0.348	0.399	0.383
	30	30	0.75	0.67	0.63	0.436	0.490	0.509	0.432	0.450	0.446
	40	40	0.81	0.72	0.69	0.390	0.453	0.472	0.392	0.427	0.420
	50	50	0.96	0.93	0.93	0.356	0.426	0.445	1.485	1.442	1.432
Fisher	SE										
	0.20	50.0	0.91	0.86	0.84	0.360	0.425	0.447	0.841	0.811	0.820
	0.25	50.0	0.85	0.77	0.74	0.357	0.425	0.446	0.340	0.383	0.401
Bayesian	0.30	48.2	0.85	0.77	0.73	0.358	0.427	0.447	0.356	0.397	0.399
	0.20	50.0	0.85	0.77	0.75	0.357	0.425	0.445	0.359	0.385	0.393
	0.25	50.0	0.85	0.77	0.75	0.356	0.425	0.446	0.341	0.375	0.390
	0.30	48.0	0.92	0.88	0.87	0.362	0.430	0.450	0.964	0.920	0.920

The bold entries represent the results of optimal MCAT designs for given conditions

traditional paper–pencil test outcomes with respect to reliability coefficients, test length, RMSD, and RMSE values, respectively.

The variation in correspondence statistics, such as RMSE and RMSD, test-length, and reliability coefficients associated with different MCAT algorithms indicate that the usage of different item selection, ability estimation, and termination rules affect the performance of the adaptive testing process. The MCAT algorithm with A-optimality and D-optimality item selection methods had similar correspondence statistics and achieved the same level of reliability when the Bayesian-MAP scoring method was utilized. However, the A-optimality item selection method showed better performance than D-optimality for both fixed test-length and precision-based stopping rules for the MCAT designs using Fisher’s scoring method.

Some studies have investigated and compared the performance of optimization-based item selection methods, which are A-optimality and D-optimality methods, in the context of MCAT (Luecht, 1996; Mulder & van der Linden, 2009; Segall, 1996). It is suggested to employ optimization-based item selection methods for the compensatory multidimensional models that allow within dimensionality at the item level where all the measured abilities are intentional (Mulder & van der Linden, 2009).

This study shows that test-length and RMSD statistics tend to decrease, while reliability coefficients tend to increase somewhat when the A-optimality item selection method is used rather than the D-optimality method. Diao and Reckase (2009) also have shown that the Bayesian-MAP ability estimation method outperforms MLE-based estimation methods for short test length. Besides, the Bayesian MAP method yields smaller RMSE values associated with each dimension for each item

Fig. 4 Random item selection MCAT results with different stopping rules

selection and stopping rule. Thus, one can suggest using Bayesian methods for ability estimation and the A-optimality method for item selection to optimize the performance of MCAT designs.

The best error variance stopping rule for each MCAT algorithm with A-optimality item selection can be considered as 0.25 for each ability estimation method with 34.6 average test length for Fisher's scoring and 27.9 average test length for the Bayesian MAP method with content balancing. Although using content balancing

with error variance criterion might lead to a somewhat increment in test length for both estimation methods, it provides more accurate and consistent results ensuring the content validity of the test.

Compared to the average test length of the English Proficiency Test (EPT), which ranges from 65 to 75, administering the EPT in MCAT format caused approximately 60% to 65% decrements in test length since the average test length of the CAT version of EPT is equal to 28 with content balancing. This finding is supported by the study conducted by Curi and Silvia (2019) in which a 25-item test was considered sufficient enough to estimate the ability scores of candidates. Similarly, a test with 25 items on average in the context of CAT has been proposed by Van der Linden and Pashley (2010).

A study conducted by Moore et al. (2018) has shown that the psychometric properties of a CAT design with 16 items are similar to the corresponding paper–pencil version of the psychological test with 74 items. This result indicates the CAT is capable of measuring the same construct with 78% fewer items compared to the corresponding paper–pencil version of it. Additionally, the CAT version of the test has higher classification accuracy compared to its shorter version with 22 items. However, other studies show inconsistency between the true-ability and estimated ability scores even when the test length was around 30 (Tseng, 2016) indicating the non-existence of consensus on the required test length. Because there are various factors affecting test length at the item and ability levels.

When it comes to the test reliability of the MCAT designs, the reliability coefficient for each dimension ranged between 0.82 and 0.95, when A-optimality was used for item selection and the Bayesian MAP method was used for the ability estimation method. However, reliability coefficients were somewhat lower for the MCAT with Fisher's ability estimation method. The reliability coefficients are suggested to be at least 0.85 if the performances of the test-takers are to be compared based on the test scores (Dai, 2015; Luo et al., 2020). The MCAT results of this study imply that the MCAT version of the EPT could measure students' language skills with high precision and validity by answering less than 40% of the items compared to the original forms.

6 Implication of this study

An important feature of the computerized adaptive testing (CAT) method is that it selects the most suitable items that match each test taker with different ability levels. As a result, the ability scores and measured trait levels are estimated more effectively and accurately with reduced test length and time (Luo et al., 2020; Pilkonis et al., 2014).

The results of post-hoc MCAT simulation based on EPT data imply that the MCAT designs have achieved higher reliability with fewer items compared to paper and pencil format. Additionally, MCAT provides ability estimates with similar reliability and a somewhat larger number of items when the content balancing is implemented ensuring the content validity. The results indicate that correspondence

statistics, test length, and reliability coefficients associated with MCAT designs are affected by the usage of different combinations of item selection, ability estimation, and stopping rules.

Among various MCAT algorithms, implementing the A-optimality item selection method, instead of D-optimality, caused a decrement in the test length and RMSD statistics, while causing a slight increment in test reliability. Therefore, it is recommended to use the A-optimality item selection method along with Bayesian- MAP ability estimation method with content balancing ensuring the content validity in the context of multidimensional adaptive testing. Moreover, it is believed that this post-hoc simulation study based on real datasets would help researchers set guidelines to develop multidimensional adaptive test versions of language tests as a strong alternative to the conventional paper–pencil testing methods.

Recently, computerized adaptive testing has gained more popularity in online learning due to its adaptive nature that allows adapting learning materials' difficulty levels (Salcedo et al., 2005) and summative assessment (Guzmán & Conejo, 2005). Additionally, CAT has been considered to be a substantial part of Massive Open Online Courses (MOOCs) (Meyer & Zhu, 2013; Oppl, et al., 2017).

The CAT versions of the psychological test have been developed as a robust alternative to their paper–pencil versions in the area of health assessment as well. For instance, CAT-NP was developed to measure narcissist personality (Luo et al., 2020), CAT-ANX for anxiety (Gibbons et al., 2014), D-CAT for depression (Fliege et al., 2005), etc. The development of these CAT versions indicates that the CAT technology has gained more popularity due to its feasibility and effectiveness with the help of advancements in technology (Luo et al., 2020).

One of the limitations of this study is that the same item bank with fixed size was used for each condition in which the total number of dimensions (or domains) are limited to 3 dimensions that are reading comprehension, listening, and grammar. Therefore, it is suggested to conduct further studies to examine the effects of item bank size, item-level dimensionality, and the number of dimensions being measured along with other factors on the test length and precision of ability estimates.

To conclude, the results of this study indicate that MCAT is feasible to measure students' multiple language skills, and it decreases the test length compared to the paper–pencil tests minimizing the burden on the test takers without compromising the precision of estimates (Lee et al., 2019; Ma et al., 2017; Ma et al., 2020). Besides, it is efficient in terms of immediate compilation of results and minimizes the possibility of cheating since each candidate takes a different set of items. Although there is a trade-off between test length and precision, it is suggested to implement content balancing to ensure content validity across different test forms and to ensure test fairness.

Acknowledgements The part of the study has been published as a Ph.D. thesis titled “Examining the Effect of Item Level Dimensionality Models on Multidimensional Computerized Adaptive Testing” and I have the approval to republish the data. The results of this study were partially presented at National Council on Measurement in Education (NCME 2016), Washington, DC, United State of America, April 9–11, 2016 .

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Availability of data and material The data was provided by the Language Department of Hacettepe University. The calibrated item parameters of EPT data were used only to construct the item bank and, to simulate the responses. The item parameters can be shared if needed. But the raw data can only be shared with approval of the Language Department.

Code availability The MAT R-package codes which are publicly available was used to simulate data and the codes in the R-packages were used to conduct the MAT analyses.

Declarations

Conflicts of interest We declare that there is no conflict of interest between the authors whatsoever.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Berger, M. P. F., & Wong, W. K. (Eds.). (2005). *Applied optimal design*. Wiley.
- Bloxom, B., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: An approximate procedure for updating*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Canada.
- Chen, S., & 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>.
- Chien, T. W., & Wang, W. C. (2017). An online multidimensional computerized adaptive testing (MCAT) module using APP. *Rasch Measurement Trans*, 31(1):1625–6. <https://www.rasch.org/rmt/rmt311.pdf>
- Curi, M., & Silvia, V. (2019). Academic English Proficiency Assessment Using a Computerized Adaptive Test. *Tendencias Em Matematica Aplicada e Computacional*, 20(2), 381–401. <https://doi.org/10.5540/tema.2019.020.02.0381>
- Dai, H.-Q. (2015). *Psychometrics*. Higher Education Press.
- Diao, Q., & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*. Retrieved [13.03.2021] from <https://www.psych.umn.edu/psylabs/CATCentral>
- Fan, M., & Hsu, Y. (1996, April). *Multidimensional computer adaptive testing*. Paper presented at the annual meeting of the American educational research association. New York City, NY.
- Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., & Rose, K. M. (2005). Development of a computer-adaptive test for depression (d-cat). *Quality of Life Research*, 14, 2277–2291. <https://doi.org/10.1007/s11136-005-6651-9>
- Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., et al. (2014). Development of the cat-anx: a computerized adaptive test for anxiety. *The American Journal of Psychiatry*, 171, 187–194. <https://doi.org/10.1176/appi.ajp.2013.13020178>

- Green, B. G., Bock, R. D., Humphries, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347–360. <https://doi.org/10.1111/j.1745-3984.1984.tb01039.x>
- Guzmán, E., & Conejo, R. (2005). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*, 48(4), 688–695. <https://doi.org/10.1109/TE.2005.854571>
- Kalender, I. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished doctoral dissertation]. Middle East Technical University
- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359–375. https://doi.org/10.1207/s15324818ame0204_6
- Kingsbury, G., & Zara, A. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests. *Applied Measurement in Education*, 4(3), 241–261. https://doi.org/10.1207/s15324818ame0403_4
- Kreitzberg, C. B., Stocking, M. L., & Swanson, L. (1978). Computerized adaptive testing: Principles and directions. *Computers & Education*, 2(4) (1978), 319–329. [https://doi.org/10.1016/0360-1315\(78\)90007-6](https://doi.org/10.1016/0360-1315(78)90007-6)
- Kullback, S. (1959). *Information theory and statistics*. Wiley.
- Lee, Y. L., Lin, K. C., & Chien, T. W. (2019). Application of a multidimensional computerized adaptive test for a Clinical Dementia Rating Scale through computer-aided techniques. *Ann Gen Psychiatry*, 18(5). <https://doi.org/10.1186/s12991-019-0228-4>
- Leung, C. K., Chang, H., & Hau, K. (2000, April). *Content balancing in stratified computerized adaptive testing designs*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.
- Lin, H. (2012). *Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model* [Unpublished Doctoral Dissertation]. The University of Illinois at Urbana-Champaign.
- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20(4), 389–404. <https://doi.org/10.1177/014662169602000406>
- Luo, H., Cai, Y., & Tu, D. (2020). Procedures to Develop a Computerized Adaptive Testing to Advance the Measurement of Narcissistic Personality. *Frontiers in Psychology*, 11, 1–12. <https://doi.org/10.3389/fpsyg.2020.01437>
- Ma, S. C., Wang, H. H., & Chien, T. W. (2017). A new technique to measure online bullying: online computerized adaptive testing. *Ann Gen Psychiatry*, 16(1). <https://doi.org/10.1186/s12991-017-0149-z>
- Ma, S., Chou, W., Chien, T., Chow, J. C., Yeh, Y., Chou, P., & Lee, H. (2020). An App for Detecting Bullying of Nurses Using Convolutional Neural Networks and Web-Based Computerized Adaptive Testing: Development and Usability Study. *JMIR Mhealth Uhealth* 8(5). <https://doi.org/10.2196/16747>
- Meyer, J. P., & Zhu, S. (2013). Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating. *Research & Practice in Assessment*, 8, 26–39.
- Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (cat) version of the schizotypal personality questionnaire. *Psychiatry Research*, 263, 250–256. <https://doi.org/10.1016/j.psychres.2018.02.022>
- Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, 74(2), 273–296. <https://doi.org/10.1007/s11336-008-9097-5>
- Oppl, S., Reisinger, F., Eckmaier, A., et al. (2017). A flexible online platform for computerized adaptive testing. *International Journal of Educational Technology in Higher Education*, 14(2), 1–21. <https://doi.org/10.1186/s41239-017-0039-0>
- Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Bulletin 69-92). Princeton, NJ: Educational Testing Service.
- Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351–356. <https://doi.org/10.1080/01621459.1975.10479871>
- Pilkonis, P. A., Yu, L., Dodds, N. E., Johnston, K. L., Maihoefer, C. C., & Lawrence, S. M. (2014). Validation of the depression item bank from the patient-reported outcomes measurement information system (promis) in a three-month observational study. *Journal of Psychiatric Research*, 56, 112–119. <https://doi.org/10.1016/j.jpsychires.2014.05.010>

- Reckase, M., D. (2009). *Multidimensional Item Response Theory: Statistics for Social and Behavioural Sciences*. New York, NY: Springer.
- Salcedo, P., Pinninghoff, M. A., & Contreras, R. (2005). Computerized adaptive tests and item response theory on a distance education platform. In J. Mira & J. R. Álvarez (Eds.), *Artificial Intelligence and Knowledge Engineering Applications: A Bioinspired Approach* (pp. 613–621). Springer.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34, 1–97. <https://doi.org/10.1007/BF03372160>
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(2), 331–354. <https://doi.org/10.1007/BF02294343>
- Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Kluwer Academic.
- Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66, 79–97. <https://doi.org/10.1007/BF02295734>
- Segall, D. O. (2004). Computerized Adaptive Testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement* (pp. 429–438). Academic.
- Sijtsma, K., & Junker, B. W. (2006). Item response theory: past performance Present Developments and Future Expectations. *Behaviormetrika*, 1, 75–102. <https://doi.org/10.2333/bhmk.33.75>
- Silvey, S. D. (1980). *Optimal design*. Chapman & Hall.
- van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, 20, 373–388. <https://doi.org/10.1177/014662169602000405>
- Song, T. (2010). *The effect of fitting a unidimensional IRT model to multidimensional data in content-balanced computerized adaptive testing* [Unpublished Doctoral Dissertation]. Michigan State University
- Sukamolson, S. (2002). *Computerized Test/Item Banking and Computerized Adaptive Testing for Teachers and Lecturers*. Information Technology and Universities in Asia – ITUA. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.211&rep=rep1&type=pdf>
- Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, 97, 69–85. <https://doi.org/10.1016/j.compedu.2016.02.018>
- van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics*, 24, 398–412. <https://doi.org/10.3102/10769986024004398>
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Springer.
- Van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). Springer.
- Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, 67(4), 575–588. <https://doi.org/10.1007/BF02295132>
- Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practices*, 12, 15–20. <https://doi.org/10.1111/j.1745-3992.1993.tb00519.x>
- Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 6, 473–492. <https://doi.org/10.1111/j.1745-3984.1987.tb00274.x>
- Wang, W. C. (2010). *Recent developments in Rasch measurement*. Paper presented at the Chair Professors: Public Lecture Series, The Hong Kong Institute of Education, China.
- Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, 35, 109–135. <https://doi.org/10.1111/j.1745-3984.1998.tb00530.x>
- Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28(5), 295–316. <https://doi.org/10.1177/0146621604265938>
- Wang, X. B., Pan, W., & Harris, V. (1999). Computerized adaptive testing simulations using real test-taker responses (LSAC Computerized Testing Report 96-06). Newtown, PA: Law School Admission Council
- Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika*, 54, 427–450. <https://doi.org/10.1007/BF02294627>
- Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. Academic Press.

- Weiss, D. J. (2005). Manual for POSTSIM: Post-hoc simulation of computerized adaptive testing. Version 2.0. St. Paul, MN: Assessment Systems Corporation.
- Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21, 361–375. <https://doi.org/10.1111/j.1745-3984.1984.tb01040.x>
- Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*, 8, 273–285. <https://doi.org/10.1177/014662168400800303>
- Yao, L. (2012). Multidimensional CAT Item Selection Methods for Domain Scores and Composite Scores: Theory and Applications. *Psychometrika*, 77, 495–523. <https://doi.org/10.1007/s11336-012-9265-5>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Burhanettin Ozdemir currently works as an Assistant Professor at the Deanship of Educational Sciences and as Director of the Center for Statistics and Information Center (CSI) in Prince Sultan University. He worked as a psychometrician in National Centre for Assessment (NCA, Qiyas) from 2017 till 2019. He received his Ph.D. degree from Hacettepe University, Ankara, Department of Educational Measurement and Evaluation in 2015. His dissertation topic is on examining the effects of different item-level dimensionality models on multidimensional computerized adaptive testing (MCAT) methods. The ultimate goal was to develop a live-CAT system that aims to measure students' language skills adaptively. He gained a scholarship from the Higher Council of Education and worked as visiting scholar at the University of Illinois at Urbana-Champaign, Illinois, during the 2014–2015 academic year under the supervision of Dr. Hua Hua Chang. His specific research interests include adaptive testing (unidimensional and multidimensional computerized adaptive testing, language testing, multistage adaptive testing), gender bias, item response theory models (IRT), structural equating modeling (SEM), multilevel modeling, multivariate statistics, test equating, DIF (differential item functioning).

Prof. Dr. Selahattin Gelbal currently works as a Professor at the Department of Educational Measurement and Evaluation and as Director of the Institute of Educational Sciences. He is also the editor of the Journal of Measurement and Evaluation in Education and Psychology. He has been teaching various psychometric courses at postgraduate levels. He has been supervising numerous master and Ph.D. dissertations. His specific research interests include computerized adaptive testing, gender inequality, item response theory models (IRT), structural equating modeling (SEM), multilevel modeling, multivariate statistics, test equating, DIF (differential item functioning).

Authors and Affiliations

Burhanettin Ozdemir¹ · Selahattin Gelbal²

Selahattin Gelbal
s.gelbal@gmail.com

- ¹ College of Humanities and Sciences, Department of Mathematics and Sciences, Prince Sultan University, Riyadh, Saudi Arabia
- ² Institute of Educational Sciences, Department of Educational Measurement and Evaluation, Hacettepe University, Ankara, Turkey