International Journal of Assessment Tools in Education
2020, Vol. 7, No. 3, 323–342
https://doi.org/10.21449/ijate.707199
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Development of a Multidimensional Computerized Adaptive Test based on the Bifactor Model

Murat Dogan Sahin[1,*], **Selahattin Gelbal**[2]

[1]Anadolu University, Department of Educational Sciences, Eskişehir, Turkey
[2]Hacettepe University, Department of Educational Sciences, Ankara, Turkey

ARTICLE HISTORY

Received: Mar. 19, 2020
Revised: July 3, 2020
Accepted: Aug. 15, 2020

KEYWORDS

Multidimensional Computerized Adaptive Testing,
Real-time Application,
Bifactor Model,
Hybrid Simulation

Abstract: The purpose of this study was to conduct a real-time multidimensional computerized adaptive test (MCAT) using data from a previous paper-pencil test (PPT) regarding the grammar and vocabulary dimensions of an end-of-term proficiency exam conducted on students in a preparatory class at a university. An item pool was established through four separate 50-item sets applied in four different semesters. The fit between unidimensional, multi-unidimensional and bifactor IRT models was compared during item calibration, with the bifactor model providing the best fit for all data sets. This was followed by a hybrid simulation for 36 conditions obtained using six item selection methods, two ability estimation methods and three termination rules. The statistics and graphs obtained indicate D-rule item selection, maximum a posteriori (MAP) ability estimation and standard error termination rule as the best algorithm for the real-time MCAT application. With the minimum number of items to be administered determined as 10, the real-time application conducted on 99 examinees yielded an average number of items of 13.4. The PPT format proficiency exam consists of 50 items, leading to the conclusion that the examinees participating in the real-time MCAT are administered an average of 74.4% fewer items than the PPT. Additionally, 86 of the examinees answered between 10-13 items. The item pool use rate is 30%. Lastly, the correlation between the PPT scores and general trait scores of 32 examinees was calculated as .77.

1. INTRODUCTION

The development of applications based on rapid and constant data flow has added momentum to studies on rapidly obtaining measurements from individuals and minimizing error levels in these measurements. To this end, it may be stated that measurement practices based on advanced technologies have gained importance from a psychometric perspective.

When measuring a trait of an individual, standard tests are commonly utilized. Due to the ease of application and to ensure understanding among individuals not versed in psychometry literature, Classical Test Theory (CTT) is frequently used for the development of these tests (Jabrayilov, Emons & Sijtsma, 2016). However, while CTT provides ease in practical application and evaluation, it carries many limitations from a psychometric perspective. It may

CONTACT: Murat Doğan ŞAHİN ✉ muratdogansahin@gmail.com Anadolu University, College of Education, Department of Educational Sciences- Eskişehir, Turkey

ISSN-e: 2148-7456 /

be stated that Item Response Theory (IRT) addresses the theoretical limitations of CTT (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). IRT posits that the estimated ability parameters are independent from the items administered to individuals. Given that test scores are equalized, this feature allows for the comparison of individuals' abilities independent from the item group (Kelecioğlu, 2001).

IRT states that just as item and ability parameters are independent from the group, standard error can be obtained for the estimated ability level of each separate individual. In addition to that characteristic, IRT also posits unidimensionality and the local independence that emerges as a result of this must be ensured to conduct scaling (van der Linden, 2016). Despite the fact that IRT is based on the assumption of unidimensionality, accepting that scales measure a single dominant latent variable contradicts the multidimensional nature of psychological constructs in practice (Reise, Morizot & Hays, 2007). Therefore, through the expansion of unidimensional IRT, multidimensional IRT emerged (Bock & Aitkin, 1981).

Due to the sophisticated mathematical foundation required by IRT, the development of the theory was stagnant until the end of the 1960's. A dominance of scientific work on IRT was observed in the 1970's (Hambeleton & Swaminathan, 1985). From this day onward, in addition to studies contributing to the theoretical development of IRT, studies were conducted comparing the ability estimations based on either CTT or IRT, obtained from the findings of tests applied to individuals. These studies indicate high correlation between IRT and CTT ability estimations for both unidimensional and multidimensional models (Gelbal, 1994; Fan, 1998; Progar & Sočan, 2008; Çelen & Aybek, 2013; Ferrando & Chico, 2007, Lawson, 1991; Ndalichako & Rogers, 1997; Akyıldız & Şahin, 2017). This situation raises the question of necessity regarding the scaling of PPT in accordance with IRT due to the complex mathematical foundations it requires. Some psychometrists posit that the purpose of IRT's existence lies in Computerized Adaptive Testing (CAT) applications (Weiss, 1985; Wainer et al., 2000; Ware et al., 2003).

Using a precalibrated item pool, CAT is an application that is based on making a provisional ability estimation for the examinee, selecting and applying the item from the pool most appropriate for the provisional ability estimation, and concluding the test in accordance with a predetermined rule (Frey, 2009; Thompson & Weiss, 2011; Bulut & Kan, 2012). A diagram of the realization of a CAT application is presented in Figure 1.

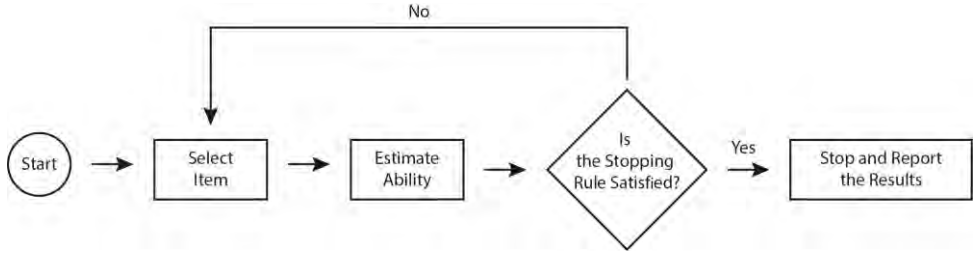

Figure 1. *CAT Applications Flow Chart*

In CAT applications; as the individuals only respond to items appropriate for their provisional ability levels, a measurement accuracy identical to a standard test that applies to the whole group is obtained through much fewer items being applied (Segall, 2005; Weiss, 2011). The ability to present individuals with items appropriate for their level in CAT applications is based on the fact that the ability level of an individual rests on the same scale as item difficulty within the scope of IRT (Reckase, 2009). Studies indicate that CAT applications provide the same measurement accuracy as PPT with 50% fewer items on average (Segall, 1996; Luecht, 1996; Eggen, 2007; Weiss & Gibbons, 2007; Gibbons et al., 2008; Weiss, 1985, 2011; Kalender & Berberoğlu, 2016).

The majority of studies on CAT applications were developed based on unidimensional IRT. However, developments in computer technologies have been increasing the interest in multidimensional CAT studies (Reckase, 2009).

Following studies in the field aiming to increase the measurement accuracy of multidimensional CAT (MCAT) compared to unidimensional CATs (e.g. Segall, 1996; Luecht, 1996), research aiming to increase the efficiency of MCAT applications grew in prominence (e.g. Veldkamp & van der Linden, 2002; Wang & Chen, 2004; Mulder & van der Linden, 2009). In the past decade, multiple studies have been conducted on developing methods regarding MCAT applications such as item selection, test termination, content balancing, etc. (Choi, Grady & Dodd, 2010; Yao, 2012, 2013, 2014; Wang, Chang & Boughton, 2012; Yao, Pommerich & Segall, 2014; Su, 2016; Lin & Chang, 2019). These studies are noted to mainly focus on within-item or between-item dimensionality. Beyond these studies, there appears to be limited research in which MCAT studies execute general trait estimation that take into account the common source of variance underlying the dimensions (sub factors) that establish the items or structure without disregarding multidimensionality (Weiss & Gibbons, 2007; Seo, 2012; Huang, Chen & Wang, 2012; Seo & Weiss, 2015; Zheng, Chang & Chang, 2013).

The purpose of this study is to portray the applicability of a PPT used to measure the grammar and vocabulary dimensions of the English proficiency of university students, following a preparatory class as an MCAT. The study consists of three main sections, in the first of which items from the proficiency exam conducted in various years as a PPT are calibrated to create an item pool. The second section consists of a hybrid simulation based on the sparse data matrix completed as a result of the missing responses created from the estimated ability levels of individuals, and the best condition for a real-time MCAT application is portrayed. The final section consists of the real-time MCAT application conducted in accordance with the algorithm based on simulation results.

In MCAT applications, multidimensional IRT models that fundamentally rely on within-item or between-item dimensionality models are used. The between-item dimensionality model (also known as multi-unidimensional model) accepts that each item measures only one dimension; however this situation is unrealistic when the nature of psychological structures are considered. The within-item model, however, assigns weight to all dimensions. In these models though, the definability of dimensions is problematic (Li & Schafer, 2005). The bifactor model used in this study provides a solution for related structures foreseen to have a general factor/ability (general trait) (Gustafson & Balke, 1993). When evaluating multidimensional constructs in order to provide the domain score, the bifactor model is considered to be highly relevant (Nieto, Abad & Olea, 2018). As such, it may be stated that this study suits the nature of English proficiency in that it will provide a general trait estimation without disregarding multidimensionality.

Thompson and Weiss (2011) state that the most important advantage of CAT applications is that they place the ability level of an individual on the same scale as item difficulty, ensuring the selection of items appropriate for the ability level of the individual being measured by the test. This ensures that individuals are only required to answer items suitable to their ability levels, resulting in a test concluded with much fewer items than they would have answered with a traditional PPT. This adaptation of the test to the individual negates the need for individuals to respond to items above or below their ability levels thereby minimizing standard error of measurement and increasing the measurement accuracy. In other words, CAT applications achieve the same measurement accuracy as traditional tests with much fewer items (Gibbons et al., 2008; Weiss, 2011). Segall (2005) states that the increase in measurement efficiency of CAT applications depends on the measurement accuracy and the length of the test, while Weiss (2011) indicates that an increase in measurement accuracy is directly related to the reduction in the number of items administered.

The fundamental components of a CAT application are; a calibrated item pool, starting rule, item selection method, ability estimation method and termination (stopping) rule (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011). Beyond these components, item exposure for the effective use of the item pool, and content balancing methods for a balanced representation of item scope may be used. However, in situations where the item pool is small, the use of item exposure dramatically increases the number of items administered to due to the limited number of items reducing the number of items equivalent to each other in terms of information function (Huebner et al., 2016). Therefore, this study does not use the item exposure method. Due to the fact that the bifactor model provides equal distribution among specific factors and their related items by default, there was no need to use any content balancing method.

2. METHOD

This study may be divided into three segments, namely calibrating the item pool, hybrid simulation, and real time MCAT application.

2.1. Item Pool Calibration

The item pool consists of 200 questions developed to measure grammar and vocabulary skills, applied at the end of a university preparatory class. Each 50 of these 200 questions were applied between 2014-2016, at the end of four different semesters. The 50 item sets were conducted on 415, 692, 798, and 1153 students in that order. During item preparation, English Language Teaching experts who have an experience of instruction and question preparation at a proficiency level contributed to the preparation, and items were prepared in accordance with the Global Scale of English (GSE) developed by Pearson.

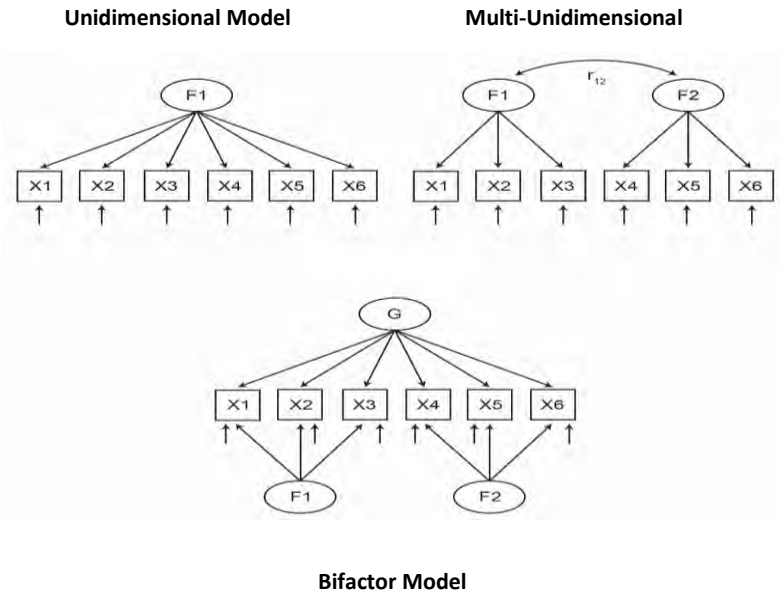

Figure 2. *Figures of the IRT Models Used in this Study*

Within the scope of this research, a multidimensional item response theory (MIRT) package (Chalmers, 2012) defined in R was used to calibrate four data sets in accordance with unidimensional, multi-unidimensional (between-item dimensionality), and bifactor models (see Figure 2). In each of these three models, 2PL was used. For each 50 item sets, a likelihood ratio chi-square statistic was used to determine whether the bifactor model improved fit over unidimensional and multi-unidimensional alternatives. It was concluded that the most appropriate approach was the bifactor model for each of the four item sets. As a result of the applications conducted to portray the invariance of the item and ability parameters, it was

observed that the the correlations between the item parameters for the lower and upper groups, and the correlations between the ability estimations determined from randomly assigning two groups of item sets were statistically significant.

2.2. Hybrid Simulation

Following the establishment of the item pool, hybrid simulation was conducted. Post-hoc simulation applications based on data obtained from the PPT application of items are used to decide the different initiation, provisional estimate of ability level, and termination rules to be used in the algorithm for the application (Weiss, 2004). During post-hoc simulations; the responses examinees provide to the items in the PPT format that establish the CAT pool are accepted as the responses they provide for the same item in the CAT application (Nydick & Weiss, 2009). Therefore, post-hoc simulations are also called "real data" simulations (Thompson & Weiss, 2011). However, the ability of a post-hoc simulation to correctly estimate a CAT output depends on all items being answered by all examinees (Weiss & Gibbons, 2007; Gibbons et al., 2008). Additionally, a complete response matrix in which examinees respond to all items cannot be obtained if item sets are applied to different groups. In such instances, completing the sparse response matrix through hybrid simulation is appropriate. Hybrid simulations use monte carlo and post-hoc simulations together to seek an answer to this question: "what would happen if all the examinees responded to all the items in the item pool?". This approach means that this question set can be tested for CAT function without the need for all items to be administered to all examinees, despite there being examinees in different groups whom have not answered some of the items in the pool.

Since the item pool in this study consists of four separate item sets applied to different groups, first, examinees' ability levels were estimated based on the 50 items they responded to, then their missing responses for the other three item sets in the sparse response matrix were generated based on their ability levels and the parameters of these items. The real and generated responses were then combined to create a 3058*200 response matrix. In turn, this matrix was used to calculate the correlation, bias, RMSD, and standard error among the θ values estimated from the PPT and hybrid simulation for 36 different conditions (see Table 1). The average number of items administered was also reported, as it is an important indicator of measurement accuracy in variable length applications. In the termination rules depending on variable test length, the minimum number of items to be administered based on the opinions of experts regarding content validity was determined as 10, while the maximum number of items in the instance that termination conditions could not be established was determined as 60. mirtCAT (Chalmers, 2016) was used for hybrid simulation applications. The initiation rule mandated by this package. It was the determination of a fixed item, therefore an item from the item pool with medium difficulty and high discrimination levels was chosen as a test initiation rule for all applications.

Table 1. *CAT Components Establishing 36 Conditions in the Simulation*

CAT Components	Method	Number of Conditions
Ability Estimation	EAP (expected a posteriori) ve MAP (maximum a posteriori)	2
Item Selaction	D-rule (the determinant rule), KL (the Kullback-Leibler divergence criteria), W-rule (weighted composite rule), weighted* W-rule, T-rule (trace of the information)and weighted* T-rule	6
Termination Rule	Standard error (.40), θ convergence ($\Delta\theta < .05$) ve fixed number of items (k=20)	3

* The weighting was determined to be for the general trait (.8, .1, .1).

2.3. Real-Time MCAT Application

In the final stage of the study, the best condition determined based on the hybrid simulation was the algorithm of the real-time MCAT application. The real-time MCAT application was conducted at the end of the preparatory class with 99 students (47 female, 52 male; age=19.3), taking advantage of the mirtCAT (Chalmers, 2016) package defined in R. For a graphical user interface (GUI), the shiny (Chang, 2019) package defined in R was used, and the researcher used their personal server during the application. An example for the interface encountered by the responder during the application is portrayed in Figure 3.

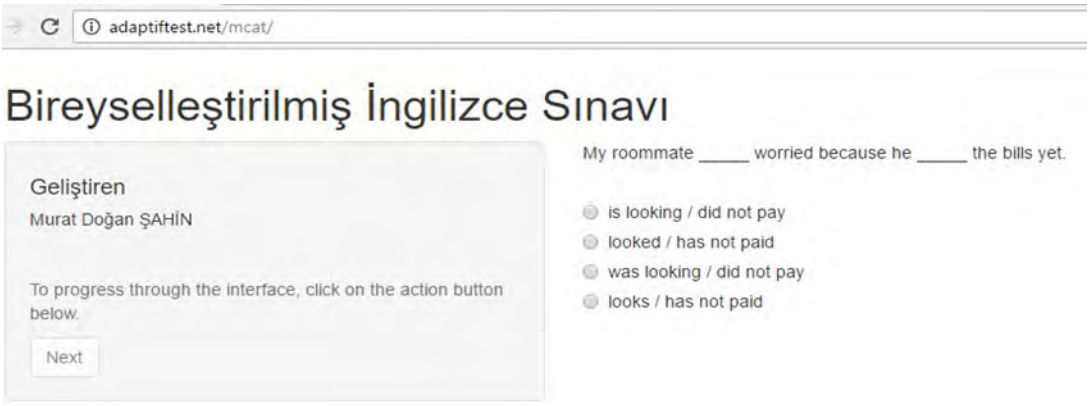

Figure 3. *GUI image of the real-time application*

It is notable that the "next" button is not active in the image above. This is due to the fact that despite not being encountered in the local application, the application enters an error state if the next button is clicked without a response to the item in the online application. As the application does not continue where it left off when this error is encountered, and a new examination application is not allowed without refreshing the server, a "javascript" applet was written to activate the "next" button when the item was responded to. The application lasted between 9-13 minutes for each student. Only one student who responded to 45 items took an 18-minute duration.

The results reflected in the database following the application show the final θ for each examinee, the standard error values for these θs, the responses for each item, the status of these responses (1-0), the ID's of the items in the database, θ and standard error histories, and lastly the time spent to respond for each item. Additionally, the correlations between the total PPT scores and the general trait scores from the real-time MCAT application of 32 students were calculated, and statistics regarding the use state of the item pool were shared.

Based on all of these practices, the research problems that emerged were as follows:

1. For the 36 different conditions within the scope of the research, taking into consideration error statistics and average number of items administered, which condition is the best for the real-time MCAT application?
2. What are the real-time MCAT application results regarding number of items administered, use rates of the item pool, and examinees' θ estimations obtained from PPT and MCAT?

3. RESULT

3.1. Hybrid Simulation Results

The results of the 36 conditions determined for the simulation were reported based on the termination rules. A study of the results obtained for the 12 conditions in which standard error termination rule is used (see Table 1) shows that under all conditions, the correlation for θ_g was

high, while the correlations for θ_2 and θ_3 were medium-low, with all being significant. While error statistics were relatively low for the general trait, they were high for specific factors. In instances where the D-rule method was used, the correlation obtained for specific factors was higher than with other methods, while the error estimations were lower. When weighting was used in item selection methods, the weighting improved the estimations obtained for θ_g as expected, while causing a drop in the values obtained for specific factors. A significant reduction in the average number of items administered (k) was observed, especially when weighting was used in the W-rule method. When the ability estimation method is being accounted for, the average number of items administered is much lower in instances using MAP compared to those using EAP. Therefore, it may be stated that MAP generally shows higher performance than EAP. The high correlations and the low standard error rates obtained for the general trait may be explained as part of the nature of bifactor structure. This is supported by the fact that one of the fundamental characteristics of the bifactor model is its explanatory power for a large portion of the variance in the variable through the general trait, while a small portion is explained by the specific factors (Reise, 2012). Therefore, it may be stated that estimations obtained for the general trait are expected to be more in line with the estimations obtained from PPTs rather than specific factors.

Regarding the faultlessness of the estimations obtained for the general trait within the framework of the standard error termination rule, all item selection methods portrayed similar performance, and weighting methods reduced the number of items administered as expected. Additionally, all other item selection methods had lower performance on specific factors compared to the D-rule method. As such, it was concluded that for the standard error termination rule, MAP ability estimation and the D-rule item selection method was the condition with the highest performance.

Following the determination of the best condition among the 12 using the standard error termination rule, the conditions based on θ convergence ($\Delta\theta < .05$) termination rule were evaluated. The results (see Table 2) obtained with MAP were found to be better than all of the item selection methods obtained with EAP. While the correlation and error statistics obtained for the general trait were similar for all the item selection methods, D-rule was found to provide the best results for specific factors once again. Regarding number of items administered, D-rule resulted in the highest values while the lowest were obtained when weighting was applied for the general trait. Despite the fact that the number of items administered to is relatively higher with the D-rule method, it portrays similar performance with other methods regarding the general trait and much better performance regarding specific factors. This led to the conclusion that the D-rule item selection method was optimal for conditions in which the θ convergence termination rule is used.

Lastly, the values obtained for the 12 conditions within the scope of the fixed number of items termination rule (k = 20) were reported (see Table 3). As with the other 24 conditions, the results show similar levels with all item selection methods of the estimated correlation and error values for the general trait in the 12 conditions where a fixed number of items termination rule is applied. The results obtained for specific factors also had high performance when the D-rule item selection method was used. The performance it provides regarding the general trait is at a similar level to other item selections and higher than them on specific factors. This resulted in the determination that use of the D-rule item selection method in conditions with a fixed number of items termination rule was more suitable, and that the results obtained with MAP were slightly better than those of EAP, concluding that this method is preferable for ability estimation.

Table 2. *Correlation, bias, RMSD, standard error values and avarage number of items administered for conditions using standard error termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE			
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	k
Standard Error (SE < .4)	EAP	D-rule	.91	.63	.66	-.0441	.0472	.0454	.38	.55	.54	.37	.69	.66	16.1
		KL	.92	.41	.58	-.0029	-.0037	-.0288	.35	.61	.55	.38	.87	.77	17.7
		W-rule	.93	.41	.58	-.0013	-.0030	.0293	.35	.61	.55	.37	.87	.78	17.8
		W-rule (.8,.1,.1)	.93	.33	.47	-.0074	.0115	.0200	.34	.62	.59	.35	.91	.83	12.9
		T-rule	.93	.44	.57	-.0273	.0249	.0476	.35	.60	.58	.35	.86	.71	13.0
		T-rule (.8,.1,.1)	.93	.35	.51	-.0081	.0187	.0328	.35	.62	.60	.34	.90	.78	12.7
	MAP	D-rule	.90	.60	.62	-.0326	.0370	.0721	.39	.56	.54	.39	.71	.65	13.4
		KL	.92	.41	.56	.0012	-.0007	.0019	.36	.61	.55	.39	.87	.78	15.2
		W-rule	.92	.40	.57	.0026	.0018	.0027	.36	.61	.55	.39	.87	.78	15.2
		W-rule (.8,.1,.1)	.92	.30	.46	.0079	.0064	.0329	.35	.62	.58	.36	.92	.84	11.4
		T-rule	.92	.42	.55	-.0195	.0084	.0806	.35	.60	.57	.36	.86	.69	11.5
		T-rule (.8,.1,.1)	.92	.33	.49	.0037	.0050	.0482	.35	.62	.58	.35	.90	.77	11.33

Table 3. *Correlation, bias, RMSD, standard error values and avarage number of items administered for conditions using ϑ convergence termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE			
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	k
θ Convergence (Δθ < .05)	EAP	D-rule	.94	.71	.71	-.0301	-.0678	.0515	.32	.53	.50	.30	.63	.62	24.1
		KL	.92	.42	.59	.0073	.0006	-.0289	.35	.60	.54	.37	.87	.77	18.1
		W-rule	.92	.41	.59	.0113	.0043	-.0282	.35	.61	.54	.37	.88	.77	18.0
		W-rule (.8,.1,.1)	.94	.39	.53	-.0031	.0173	.0152	.32	.62	.57	.31	.88	.79	16.1
		T-rule	.94	.50	.62	-.0164	.0230	.0473	.32	.59	.55	.30	.80	.68	17.1
		T-rule (.8,.1,.1)	.94	.41	.55	-.0095	.0216	.0337	.32	.62	.58	.30	.86	.75	16.3
	MAP	D-rule	.94	.70	.70	-.0069	.0639	.0768	.32	.52	.49	.33	.63	.61	21.4
		KL	.92	.41	.59	.0208	.0013	-.0072	.36	.60	.54	.38	.87	.76	17.2
		W-rule	.92	.40	.59	.0218	.0037	-.0032	.36	.61	.54	.38	.87	.76	17.1
		W-rule (.8,.1,.1)	.94	.36	.53	.0108	.0078	.0320	.32	.62	.56	.32	.88	.77	15.1
		T-rule	.94	.50	.63	.0030	.0086	.0684	.32	.59	.53	.31	.79	.65	16.6
		T-rule (.8,.1,.1)	.94	.42	.53	.0088	.0055	.0570	.32	.61	.57	.32	.85	.74	15.0

Table 4. *Correlation, bias, RMSD and standard error values for conditions using fixed number of items termination rule*

Termination Rule	Ability Estimation Method	Item Selection Method	r			bias			RMSD			SE		
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$
Fixed Number of Items (k = 20)	EAP	D-rule	.93	.69	.69	-.0315	.0637	.0481	.33	.53	.51	.34	.65	.63
		KL	.93	.42	.60	.0085	-.0023	-.0344	.33	.61	.54	.37	.87	.76
		W-rule	.93	.42	.61	.0108	-.0012	-.0336	.33	.61	.53	.35	.87	76
		W-rule (.8,.1,.1)	94	.42	.60	-.0093	.0183	.0115	.30	.62	.55	.29	.85	.72
		T-rule	.95	.53	.66	-.0154	.0277	.0346	.30	.58	.53	.29	.78	.66
		T-rule (.8,.1,.1)	.95	.50	.58	-.0097	.0231	.0266	.30	.59	.57	.28	.81	.72
	MAP	D-rule	.93	.69	.69	-.0093	.0607	.0751	.33	.52	.50	.33	.64	.61
		KL	.93	.43	.61	.0170	-.0054	-.0086	.33	.60	.53	.35	.86	.75
		W-rule	.93	.42	.61	.0184	.0009	-.0061	.33	.61	.53	.35	.86	.75
		W-rule (.8,.1,.1)	.95	.42	.61	.0078	.0071	.0255	.29	.62	.53	.30	.84	.70
		T-rule	.95	.53	.66	-.0027	.0144	.0601	.29	.58	.52	.29	.76	.64
		T-rule (.8,.1,.1)	.95	.50	.58	.0037	.0124	.0495	.29	.59	.55	.29	.80	.80

To determine the most suitable condition for use in the real-time MCAT as a result of the simulations, a final evaluation was conducted for the three conditions with the best results for all termination rules. The error rates and correlation statistics of these conditions are provided in unison (see Table 4), and the lavaan (Sarkar, 2016) package in R was used to graph each one individually (see Figure 4), with the best condition for the real-time MCAT application being decided as a result of these values and graphs.

For three different termination rules, the best results were obtained using D-rule item selection and MAP ability estimation methods. Of these three conditions, the one with the highest measurement accuracy for the real-time application was determined by studying the graphs obtained for the general trait.

Table 5. *Statistics of the best conditions for each termination rule*

Termination Rule	Ability Est. Method	Item Selection Method	r			bias			RMSD			SE			k
			$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	$\theta_{(g)}$	$\theta_{(2)}$	$\theta_{(3)}$	
Standard Error	MAP	D-rule	.90	.60	.62	-.033	.037	.072	.39	.56	.54	.39	.71	.65	13.4
θ Convergence	MAP	D-rule	.94	.70	.70	-.007	.064	.077	.32	.52	.49	.33	.63	.61	21.4
Fixed Number	MAP	D-rule	.93	.69	.69	-.009	.061	.075	.33	.52	.50	.33	.64	.61	20

Firstly, the standard error - θ_g graph for the three conditions was obtained for the general trait. In this case, as termination is based on .4 standard error, despite only the maximum (60) number of items administered, the estimations that don't fall below this standard error value are still above .4. It is notable that these high standard error values are observed with individuals with high θ levels.

The second and third graphs were obtained for θ convergence and for fixed number of items termination rules, and these graphs appear similar to 3 each other. Both graphs have a very small range for standard error values towards the center of the ability scale. However, as the estimated θ value of examinees increases, the standard error value increases and the values obtained go as high as .6. This situation may stem from the fact that the medium level ability estimations (θ = 0) of the item pool provide more information, while anything beyond θ = 1 provides less information. It is also notable that the standard error value obtained with θ convergence disperses over a wider range compared to that obtained with fixed number convergence.

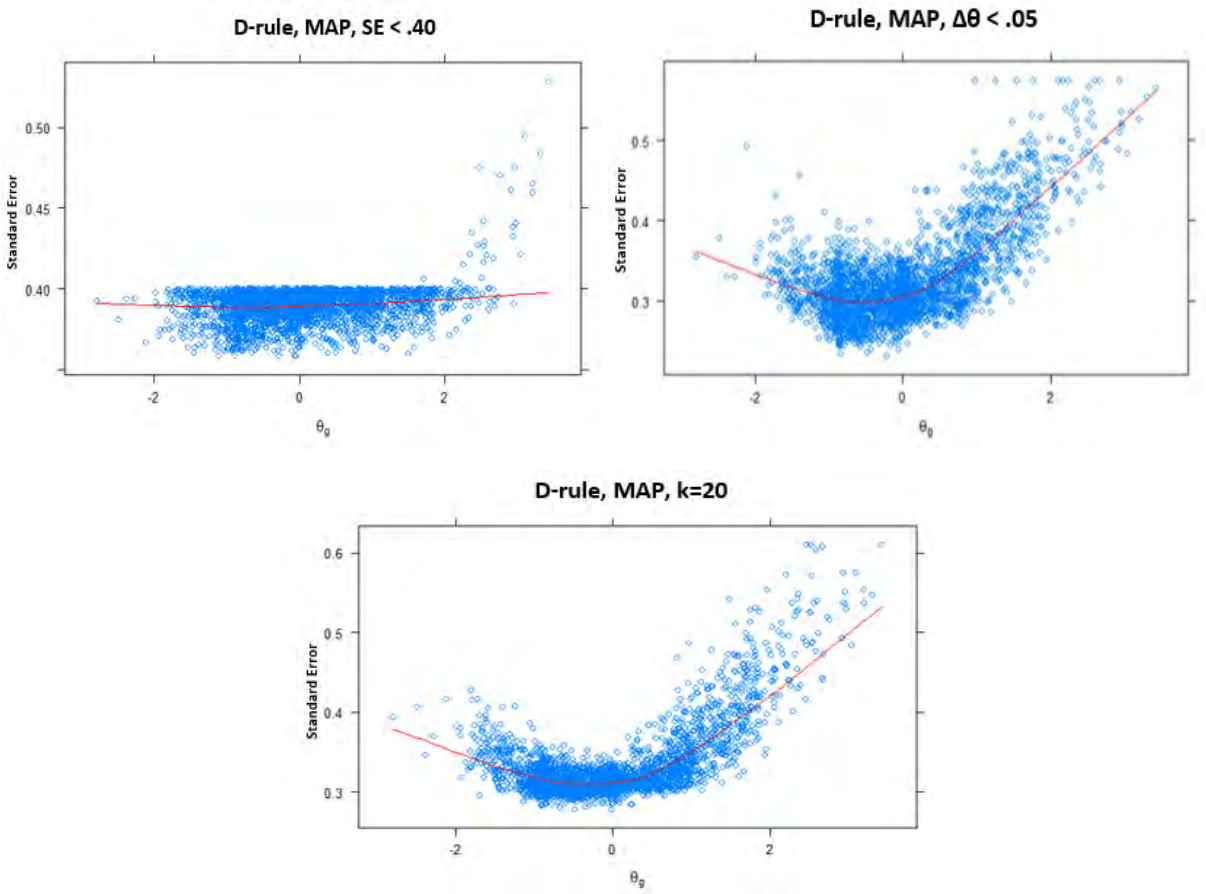

Figure 4. *Standard error – θ_g graphs*

When the number of items administered – θ_g graphs obtained for variable length applications are studied (see Figure 5), in cases where the standard error termination rule is used; the average number of items administered is near 10 throughout a large portion of the θ scale, and this value increases as θ approaches 2. It is observed that individuals with high ability levels reached the maximum number of items to be administered, in addition to the termination rule. In the condition where the θ convergence termination rule is used, it is notable that the average number of items administered over the whole ability scale has a high and wide range.

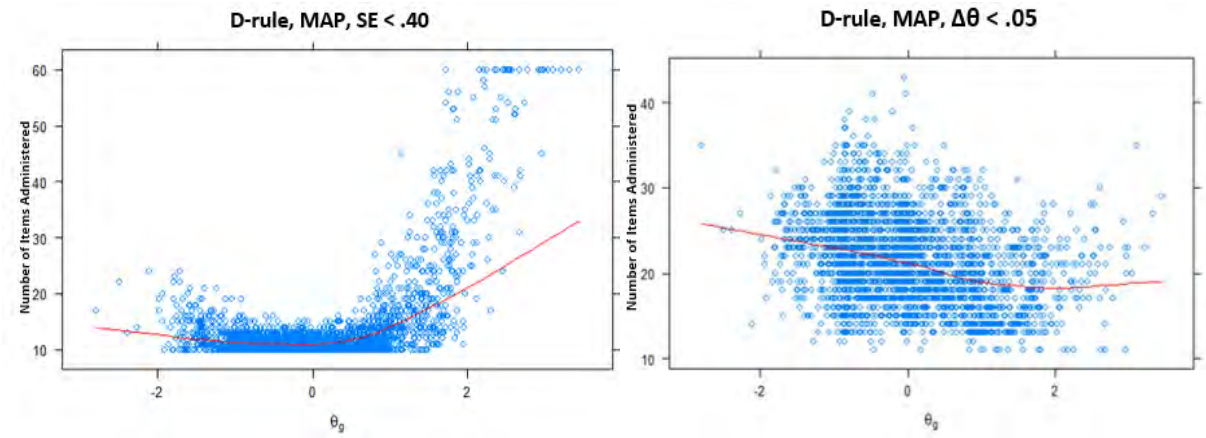

Figure 5. *Number of items administered - θ_g graphs*

In conditions where standard error termination rule in the hybrid simulation are used and the frequency values of the item numbers responded are studied (see Figure 6), it is notable that approximately 30% of the 3057 participants responded to 10 items, the minimum determined to terminate the test. Additionally, based on this graph, it may be stated that approximately 85% of the individuals responded to 10-15 items.

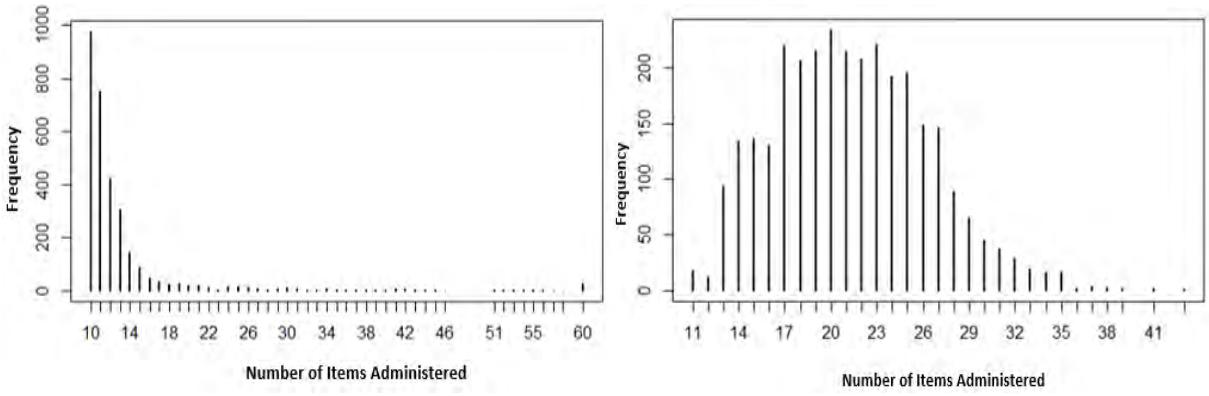

Figure 6. *Frequency of number of items administered for standard error & θ convergence termination Rules*

In a large number of participants, the number of items administered in the condition using the θ convergence termination rule varied between 17-25. The difference from the condition using the standard error termination rule is that the range of the number of items answered in this condition is narrower.

Based on these graphs, it may be stated that the use of the standard error termination rule in the real-time application is more efficient than other methods regarding number of items administered. After a comparison of the graphs and statistics obtained for the best conditions, as a result of the simulations for each termination rule; MAP was selected as the ability estimation method, D-rule was selected as the item selection method, and standard error as a termination rule (.4) was selected as the most appropriate components of the algorithm for the real-time MCAT application.

3.2. Real-Time MCAT Application Results

Based on the results obtained for 36 different conditions regarding the hybrid simulation, D-rule was chosen as the item selection method, MAP as the ability estimation method, and a .40 value cutoff in the standard error for the general trait as a termination rule was decided on during the real-time MCAT application. In addition, a minimum of 10 items to be administered to each participant for the test termination, and test termination after 60 items in instances where standard error remained above .40 criteria were applied. Based on this algorithm, the real-time MCAT application was conducted using the mirtCAT (Chalmers, 2016) package and the shiny (Chang, 2019) GUI package for R on 99 students in the final semester of the preparatory class.

Studying the frequencies of the number of items administered (see Table 6) shows that 74 participants answered 10-12 items. 12 participants answered 13 items, while the number of participants who responded to 14 items was 4, and 15 items was 5. Only 4 participants answered more than 15 items.

The results obtained show that the average number of items administered to the 99 students participating in the real-time application is 12.3. This value is close to but slightly lower than the average number of items of 13.4 obtained during the simulation application using the same condition (D-rule, MAP, SE<.4) as the real-time application. The number of items administered varies between 10 and 45. Regarding the grammar and vocabulary skills measured by the real-

time application, the number of items examinees answered in the PPT is 50. This led to the conclusion that in the real-time MCAT application, examinees are administered an average of 74.4% fewer items than the PPT.

Table 6. *Distribution of number of items administered during the real-time application*

Number of Items Administered	Frequency	%
10	25	25.3
11	24	24.2
12	25	25.3
13	12	12.1
14	4	4.0
15	5	5.1
>15	4	4.0
Total	99	100.0

Following the real-time MCAT application, it was observed that 60 of the items in the 200 present in the item pool were used, while 140 were not present in any of the applications. In other words, in the real-time MCAT application conducted with 99 individuals, 30% of the item pool was used. Of the 60 items used, it is notable that 37 of them have used numbers under 5.

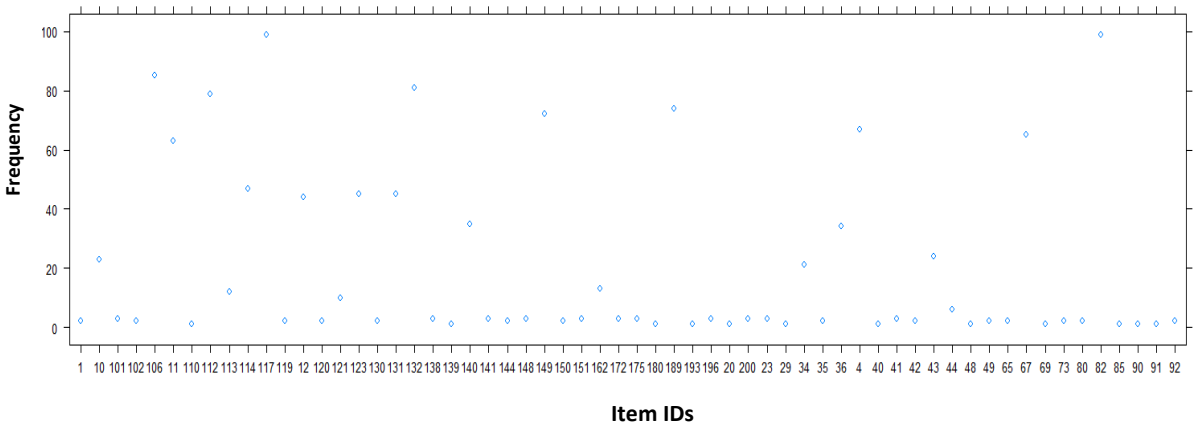

Figure 7. *Item use from the item pool*

When the use frequencies of the 60 items from the item pool used at least once in the real-time application are studied (see Figure 7), it was found that of these items, item number 117 was used at the beginning of every application, and the 82nd item was present in all the applications. Other than these two items, 8 items were administered at least in 60 applications. The number of items administered 20 or more times was 19.

The real-time MCAT application was conducted on 99 students studying at an English preparatory class at a university in Turkey. Of these students, 32 entered their proficiency examination one month before the application was conducted. Through this opportunity, the correlation between the real-time MCAT application and their PPT results (proficiency examination) were calculated. These calculations resulted in a .77 correlation between the general trait estimations resulting from the real-time MCAT application and their total score obtained from the PPT.

4. DISCUSSION and CONCLUSION

Within the scope of this study, grammar and vocabulary data of English preparatory class students were gathered from their proficiency examinations required to attend undergraduate

courses, and an MCAT measuring the general trait and their grammar and vocabulary was developed. To this end, an item pool consisting of four separate groups with 50 items each was established. This was followed by a hybrid simulation application to determine the algorithm to be used in the real-time MCAT application. Following this simulation, the ability estimation methods (EAP and MAP), item selection methods (D-rule, KL, W-rule, T-rule, weighted W-rule and weighted T-rule), and the termination rules (standard error, θ convergence and fixed number of items) were used to create 36 different conditions. For each dimension in these conditions, the correlation between the real and estimated θ values, bias, RMSD and standard error values were obtained. Due to the fact that in addition to correlation values and error statistics, the average number of items administered is also an important indicator of measurement accuracy in CAT applications, the number of items administered in the conditions using termination rules based on variable test length were also reported. Following the determination of the most appropriate MCAT algorithm based on the simulations for the real-time application, this algorithm was used to conduct the real-time MCAT application. The correlation between the PPT scores and MCAT real time application results of the 32 examinees was also calculated. Additionally, the item use frequencies of items in the pool and number of items administered to each 99 examinee participating in the application were reported. In this section, the results are presented under separate headings for the hybrid simulation and the real-time MCAT application.

4.1. Interpretation of Findings Obtained from the Hybrid Simulation

The simulation results indicate that for the three termination rules used within the scope of the study, the most appropriate conditions the ones where D-rule item selection and MAP ability estimation methods used. While the D-rule item selection method provided similar performance with other methods regarding general ability estimation, it provided much better values than other methods for specific factors. These findings are similar to the study of Seo and Weiss (2015), who suggested the use of D-rule item selection and MAP ability estimation methods in situations where estimations for specific factors are important.

Within the scope of this study, the correlations between the real and estimated ability parameters for the general ability were quite high under all conditions, while the correlations for specific factors were lower. The first reason for this may be the nature of the bifactor model. This is due to the known given that for a multidimensional model to fit with a bifactor structure, the structure must not only estimate a general ability but the factor loadings for the general ability must be higher than group factors (Reise, Morizot & Hays, 2007). It may be stated that this situation causes a reduction in given information for specific factors as a structure adapts to the bifactor model. In addition, Seo (2011) found that similar to this study, the correlation values obtained for the general ability are higher than those obtained for specific factors.

While the correlation values obtained for the general ability were high, another reason the values for group factors being low is the number of items administered. Weiss and Gibbons (2007) indicate that to increase the efficiency of bifactor MCATs, between 20 and 50 items must be used for each specific factor. Other related studies in the literature on bifactor MCAT applications such as Seo (2011) and Seo and Weiss (2015) also used 20 items for each group factor. In Sunderland et al.'s (2019) study, which aimed to estimate internalizing through a bifactor MCAT application, it was found that a 133 item PPT scale was completed in an average of 44 items. Nieto, Abad and Olea (2018) developed an MCAT application based on a bifactor model of the big five scale, and concluded that a result was obtained for each dimension through 12 items on average. Within the scope of this study, 10 items were used for each factor in the fixed number of items termination rule condition, while the number of items answered fell as low as 5 for each specific factor for a large portion of the other conditions. This may be the cause of the low correlations obtained for specific factors and the high error statistics. Within

the scope of this study, the researcher aimed to develop a real-time application for a 50-item PPT application. The number of items administered in CAT studies is directly related to measurement accuracy. Additionally, considering the real-time application of this study aims to obtain an overall score estimation without disregarding multidimensionality, it was predicted that determining the minimum number of items to be answered for each dimension as 20 would reduce the efficiency of the real-time application.

When the item selection methods with weighting were studied, within termination rules based on variable test length, use of W-Rule item selection methods with weighting results in a rise in the correlations for general ability and a significant reduction in error statistics. The number of items answered with weighting was reduced by 20-25% on average. Additionally, the improvement in the performance of T-rule with weighting was higher than with W-rule.

In applications using the standard error termination rule, especially with ability levels where the item pool information level is low, the estimated standard error levels were observed to be high. As such, in applications where the item pool is not large enough, it may be stated that the use of a standard error termination rule is more appropriate.

4.2. Interpretation of Real-Time MCAT Application Findings

Following the real-time MCAT application, 30% of the 200 item pool was used. Based on these values, it may be stated that the use rate of the pool is low. However, studies indicate that in 50% of CAT applications, only 14% of the item pool is used (Wainer, 2000). Considering the item pool information level is high for a mid-low θ level, middle or low ability levels of the examinees may be causing the use of only a small portion of the pool. In addition, as the students who participated in the real-time MCAT application are from the same course level and therefore at similar ability levels regarding the test, the high use rate of certain items is to be expected. In such instances, the use of very easy and very difficult items are expected to be low (Wei & Lin, 2015). Additionally, the average number of items administered being low at 12.3 and the lack of an item exposure control method within the scope of the study may also be causes behind the low use rates of the item pool. The generally similar ability levels and the lack of an item exposure rate control mechanism leads to the conclusion that the limited number of items administered are frequently the same items.

Of the 60 items used in real-time MCAT application, only 23 were included in 5 or more of the 99 applications. This situation is similar to Veldkamp and van der Linden's (2002) MCAT study in which over 80% of the tests only used 20% of the item pool. As stated earlier, as the average number of the items administered is low and 75% of the examinees responding to fewer than the average number of items administered may have been effective in this situation emerging.

The findings show that the real-time MCAT application lasts approximately 9 minutes. When considering each question is allocated one minute in a PPT, it may be stated that the MCAT application takes 80% less time than PPT. This is considered to be important regarding the effectiveness of CAT applications.

4.3. Recommendations for Future Research

Within the scope of this study, while conducting a general ability estimation based on a common source of variance for all items, a bifactor model that takes into account multidimensionality was used. It may be stated that bifactor models are an alternative to high-order/hierarchical models (Seo & Weiss, 2015). The only MCAT study in the literature using high-order IRT models was conducted by Huang, Chen and Wang (2012). It is believed that a comparison between the findings of this current study and an MCAT study using high order IRT models would contribute significantly to the literature.

Despite the increase in the number of studies in the recent years on bifactor MCAT, the literature in this field is still limited. Some of these studies were conducted beyond the scope of the purpose of this study (Zheng et al., 2013; Gibbons et al., 2016). It is also noted that all of the applications conducted regarding real-time applications in bifactor MCAT studies aim to study affective characteristics. It may therefore be stated that a need has arisen for bifactor MCAT applications with different situations in which the aim is to produce an overall score for a multidimensional cognitive ability – as with this study.

The item pool used within the scope of this study is limited. Future research may use item exposure control methods in real-time applications based on larger item pools, and these methods may allow a comparison between the performance of item selection and ability estimation methods.

Within this study, only items with independent response status were used. However, the measurement of language skills also requires applications based on the response to more than one item based on a text, image, etc. As such, it is documented that testlet-based IRT models may be used (e.g. Frey, Seitz & Brandt, 2016). Additionally, the bifactor model used within the scope of this study may be used for testlet-based tests (see DeMars, 2006). Thus, it is believed that an MCAT application based on the bifactor model for tests of skills beyond the scope of this study such as reading and listening, which are some of the fundamental dimensions of language skills, would contribute to the research.

Acknowledgements

This study presents partial findings of the doctoral dissertation entitiled *Examining the Results of Multidimensional Computerized Adaptive Testing Applications in Real and Generated Data Sets* (Şahin, 2017).

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

ORCID

Murat Doğan ŞAHİN https://orcid.org/0000-0002-2174-8443
Selahattin GELBAL https://orcid.org/0000-0001-5181-7262

5. REFERENCES

Akyıldız, M. & Şahin, M. D. (2017). Açıköğretimde kullanılan sınavlardan Klasik Test Kuramına ve Madde Tepki Kuramına göre elde edilen yetenek ölçülerinin karşılaştırılması. *AUAd, 3*(4), 141-159.

Bulut, O & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49*, 61-80.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algortihm. *Pschometrika, 46*(4), 443-459.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software, 71*(5), 1-38.

Chang, W. (2019). Shiny: Web application framework for R. Version 1.3.2

Choi, S. W., Grady, M. W., & Dodd, B. G. (2010). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement, 71*, 37-53.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement. 43*(2), 145–168.

Eggen, T. (2007). *Choices in CAT models in the context of educational testing*. Paper presented at the CAT Models and Monitoring Paper Session, June 7, 2007 (Retrieved November 11, 2016, from http://publicdocs.iacat.org/cat2010/cat07eggen.pdf).

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

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-374.

Ferrando, P. & Chico, E. (2007). The external validity of scores based on the twoparameter logistic model: Some comparisons between IRT and CTT. *Psicológica, 28*, 237-257.

Frey, A. & Nicki-Nils, S. (2009). Multidimensional adaptive testing in educational and psychological measurement: Current state and future challenges. *Studies in Educational Evaluation, 35*. 89-94

Frey A, Seitz N-N and Brandt S (2016) Testlet-Based Multidimensional Adaptive Testing. *Front. Psychol., 7*, 1758.

Gelbal, S. (1994). *P madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma*. Unpublished doctoral dissertation. Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual Review of Clinical Psychology, 12*, 83-104.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4), 49-58.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407-434.

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

Huang, H., Chen, P & Wang, W. (2012). Computerized adaptive testing using a class of high-order item response theory. *Applied Psychological Measurement, 36*(8), 689-706.

Huebner, A. R., Wang, C., Quinlan, K. & Seuber, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behav. Res., 48*, 1443-1453

Jabrayilov, R., Emons, W. H. M. & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement, 40*(8) 559-572.

Kalender, I., & Berberoglu, G. (2017). Can computerized adaptive testing work in students' admission to higher education programs in Turkey? *Educational Sciences: Theory & Practice, 17*, 573-596.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki, *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20*, 104-110.

Lawson, S. (1991). *One parameter latent trait measurement: Do the results justify the effort?* In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments. Greenwich, CT: JAI.

Li, Y. H., & Schafer, W. D. (2005). Trait parameter recovery using multidimensional computerized adaptive testing in reading and mathematics. *Applied Psychological Measurement, 29*, 3-25.

Lin, C. & Chang, H. (2018). Item Selection Criteria with Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing. *Educational and Psychological Measurement, 79*(2), 335-357.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*(4), 389-404.

Ndalichako, J. L., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement, 57,* 580-589.

Nieto, M. D., Abad, F. J., & Olea, J. (2018). Assessing the Big Five with bifactor computerized adaptive testing. *Psychological Assessment, 30*(12), 1678-1690.

Nydick, S. & Weiss, D. J. (2009). *A hybrid simulation procedure for developments of CATs*. Paper presented at the Item Pool Development Paper session at the 2009 GMAC Conference on Computerized Adaptive Testing.

Progar, S. & Sočan, G. (2008). An empirical comparison of Item Response Theory and Classical Test Theory. *Horizons of Psychology, 17*(3), 5-24.

Reckase, M., D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. New York, NY: Springer.

Reise, S. P. (2012). The rediscovery of bifactor measurement models, *Multivariate Behavioral Research, 47*(5), 667-696.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*, 19-31.

Sarkar, D. (2016). *Lattice: Multivariate Data Visualization with R*. Springer.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

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

Seo, D. G. (2011). *Application of the bifactor model to computerized adaptive testing*. Unpublished Doctoral Disertation. University of Minnesota.

Seo, D. G. & Weiss, D. J. (2015). Best Design for Multidimensional Adaptive Testing with the Bifactor Model. *Educational and Psychological Measurement, 75*(6), 954-978.

Su, Y. (2016). A comparison of constrained item selection methods in multidimensional computerized adaptive testing. *Applied Psychological Measurement, 40*(5) 346-360.

Sunderland, M., Batterham, P. Carragher, N., Calear, A. & Slade, T. (2019). Developing and Validating a Computerized Adaptive Test to Measure Broad and Specific Factors of Internalizing in Community Sample. *Assessment, 26*(6) 1030-1045.

Şahin, M. D. (2017). *Examining the Results of Multidimensional Computerized Adaptive Testing Applications in Real and Generated Data Sets* [Unpublished doctoral dissertation]. Hacettepe University, Graduate School of Educational Sciences, Ankara.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9.

van der Linden, W. J. (2016). *Handbook of Item Resonse Theory*. Boca Raton: CRC Press.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575-588.

Wang, C., Chang, H. & Boughton, K. A. (2012). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99-122.

Wainer, H. W., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25,* 203-224.

Ware J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlo, C. G. H., Tepper, S. & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935-952.

Wei, H., & Lin, J. (2015). Using out-of-level items in computerized adaptive testing. *International Journal of Testing, 15*(1), 50-70.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70-84.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences, 2*(1), 1-27.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of Computerized Adaptive Testing to Educational Problems. *Journal of Educational Measuremen, 21*(4), 361-375.

Weiss, D. J., & Gibbons, R. D. (2007). *Computerized adaptive testing with the bifactor model*. Paper presented at the New CAT Models session at the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved-October 12, 2016, from http://publicdocs.iacat.org/cat2010/cat07weiss&gibbons.pdf

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and Applications. *Psychometrika, 77*, 495-523.

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*, 3-23.

Yao, L. (2014). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constraints. *Journal of Educational Measurement, 51*, 18-38.

Yao, L., Pommerich, M & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet price, screening test. *Applied Psychological Measurement, 38*(8) 614-631.

Zheng, Y., Chang, C. H., & Chang, H. H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 22*, 491-499.