ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(4);422-435

Can TIMSS Mathematics Assessments be Implemented as Computerized Adaptive Test?

Semirhan GÖKÇE* Cees A.W. GLAS**

Abstract

In recent years, there has been a growing interest and extensive use of computerized adaptive testing (CAT) especially in large-scale assessments. Numerous simulation studies have been conducted on both real and simulated data sets to determine the optimum conditions and develop CAT versions. Being one of the most popular large-scale assessment programs, Trends in International Mathematics and Science Study (TIMSS) has been implemented as paper and pencil tests to monitor student achievement in mathematics and science at fourth and eighth grade levels since 1995. The purpose of this study is to investigate the optimum CAT algorithm for TIMSS eighth grade mathematics assessments. Since Turkey and USA participated in 2007, 2011 and 2015 administrations, their data were combined and then 393 items were calibrated on the same scale by using marginal maximum likelihood estimation method. With this item pool, several scenarios were proposed and tested to determine not only the optimum starting rule, ability estimation method, test termination rule but also the efficiency of exposure control method. The results of the study indicated that estimating abilities with expected a posteriori method after 6 random items, terminating the fixed-length test after 20 items seemed to be the optimum algorithm for TIMSS eighth grade mathematics assessments. Also, it was found that using item exposure control had a prior importance for the effective use of the item pool. This study has some implications for both national and international large-scale test developers in determining the optimum CAT algorithm and its consequences compared with paper and pencil versions.

Key Words: computerized adaptive testing, item response theory, mathematics assessment, simulation study, TIMSS.

INTRODUCTION

Educational testing has mainly been focused on traditional paper and pencil tests until the technological developments have supported the emergence of computers. At first, computers were responsible for displaying items and collecting responses, but since then they have also supported innovative item formats (Zenisky & Sireci, 2002) and fast score reporting. Then, instead of administering same set of items to the participants, different test forms have been assembled in computer-based testing. Eventually, this becomes meaningful when the participant's cumulative performance on earlier items determines the selection of newer items (Davey & Pitoniak, 2006). Actually, this is the main idea behind computerized adaptive testing (CAT). The intuitive principle underlying CAT is to maximize the item information. Statistically, each item gives information about the participants in terms of the trait being measured, but when the item parameters fit to their interim ability estimations, the amount of information maximizes. Therefore, the correct response of a participant is followed by more difficult item and the incorrect response is followed by an easier item (Hambleton, Swaminathan, & Rogers, 1991; Luecht & Sireci, 2012; van der Linden, 2010). This optimization process continues until the test administrators have enough certainty about the sufficiency of information about participant's ability level. Unlike traditional tests in which all participants take a single form, the CAT algorithm tailors the items according to the response patterns (Sireci, Baldwin, Martone, Kaira, Lam, & Hambleton, 2008) and finitely many test forms can be created during test

* Asst. Prof. Dr., Niğde Ömer Halisdemir University, Niğde-Turkey, e-mail: semirhan@gmail.com, ORCID ID: 0000-0002-4752-5598
** Prof. Dr., University of Twente, Enschede-The Netherlands, e-mail: c.a.w.glas@utwente.nl, ORCID ID: 0000-0001-6531-5503

To cite this article:
Gökçe, S., & Glas, C. A. W. (2018). Can TIMSS mathematics assessments be implemented as computerized adaptive test?. *Journal of Measurement and Evaluation in Education and Psychology, 9*(4), 422-435. DOI: 10.21031/epod.487351
Received: 25.11.2018
Accepted: 21.12.2018

administration. In this manner, different types of computer based tests range in a wide spectrum, from linear tests to adaptive tests.

Compared with linear tests in which fixed test forms are used, CAT has many advantages, such as testing on demand (Glas & Geerlings, 2009; Hambleton et al., 1991; van der Linden, 2001; Wainer, 2000), shortening tests without loss of measurement precision (Eggen, 2007; Hambleton et al., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010), enabling immediate test scoring (Eggen, 2007; Wainer, 2000) and minimizing test frustration (Hambleton et al., 1991, Mills & Stocking, 1996). On the other hand, CAT has also some disadvantages such as reducing the control over tests and requiring a large calibrated item bank (Meijer & Nering, 1999).

The theoretical framework of CAT is based on Item Response Theory (IRT) framework in which the probability of a correct response to an item can be written as a mathematical function of participant's ability and item parameters. With IRT, the ability estimations of the participants can be obtained by independent set of items administered with a standard error. Hambleton et al. (1991) states that IRT provides a framework for comparing the ability estimations of differrent participants even if they have different set of items. Therefore, in order to match the item parameters with the ability levels of the participants, a large set of items (it is called item pool or item bank) is required whose statistical characteristics are obtained. van der Linden (1995) lists four steps of developing an iterative CAT algorithm as (1) defining the starting rule, (2) deciding on the item selection criteria, (3) choosing the ability estimation method, and (4) determining the termination rule. While determining the optimum starting rule, the difficulty of the first few items is important. Many testing programs have been using easier items at the beginning of a test in order to provide an initial success experience or motivation of the participants (Mills & Stocking, 1996). As the item selection methods are concerned, mainly there are two approaches such as Fisher's maximum information and Bayesian methods. Although Wainer (2000) states that both of the item selection methods give good results, Bayesian criteria needs more demand on the computer capabilities (Eggen, 2004). As the ability estimation methods are discussed, there are four ability estimation methods: maximum likelihood (ML), weighted maximum likelihood (WML), maximum a posteriori (MAP) estimation and expected a posteriori (EAP) estimation. According to Gu and Reckase (2007), MAP and EAP produce smaller standard errors compared to MLE and WML for the same number of items but they may produce biased estimations for inappropriate prior distributions. In test termination, there are mainly two options either to use fixed-length test or variable-length test. The former guarantees the implementation of a specified number of items to each participant but ends up with different standard error values for ability estimations. On the other hand, the latter stops the algorithm either obtaining sufficiently accurate ability estimation by comparing the standard error with a reference value or looking at the difference between consecutive ability estimations. At this point, the test developers should decide on test termination rule either to use a fixed-length test or a variable-length test depending on the purpose of the test and the content validity as well.

Due to the development of information and communication technologies and the widespread use of computers, many large-scale tests have been implemented as computer based test or even CAT such as Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT), Armed Services Vocational Aptitude Battery (ASVAB), and United States Medical Licensing Examination (USMLE). GRE, which was developed by Educational Testing Service (ETS), was implemented as a CAT as of 1992, Graduate Management Admission Council's GMAT was implemented as a CAT as of 1997 (Luecht & Sireci, 2012).

Trends in Mathematics and Science Study (TIMSS) is also a large scale assessment program aimed to monitor student achievement in mathematics and science at fourth and eighth grade levels in four-year-cycle since 1995 (Mullis, Martin, & Loveless, 2016). TIMSS assessments have been administered in paper-and-pencil form and the achievement tests have 14 different booklets which are linked to each other by common items, i.e. anchor items. In the booklets, there are both multiple-choice and open-ended items. Also, there have been anchor items between any consecutive TIMSS assessments so that test equating becomes feasible across assessments.

Purpose of the Study

The purpose of this study is to investigate the optimum CAT algorithm alternative to the paper and pencil based TIMSS eighth grade mathematics assessments. The data of two participating countries in the TIMSS eighth grade mathematics assessments in 2007, 2011 and 2015, Turkey and USA, were used for item calibration and the item pool. Then, a series of simulations covering different scenarios were tested to compare different starting rules, ability estimation methods and test termination rules. Additionally, item exposure rates were calculated in order to determine the effect of item exposure control strategy. The research questions of the study are given below.

1. What is the optimum CAT algorithm of TIMSS eighth grade mathematics assessments regarding different starting rules, ability estimation methods and test termination rules?

2. How does the item exposure control strategy affect the optimum CAT algorithm which is developed as an alternative to TIMSS eighth grade mathematics assessments?

METHOD

This part contains information related with the participants, data collection instruments and data analysis.

Participants

Table 1 gives information about the TIMSS sample sizes of Turkey and United States of America in 2007, 2011 and 2015 eighth grade mathematics administrations.

Table 1. Participants of Turkey and USA in TIMSS eighth grade mathematics assessments

Year	Turkey	USA	Total
2007	4498	7377	11875
2011	6928	10477	17405
2015	6079	10221	16300
Total	17505	28075	45580

Data Collection Instruments

As mentioned before, 14 different booklets were used in TIMSS eighth grade mathematics assessments and these booklets were linked to eachother with anchor items. Table 2 shows the number of items in these booklets.

Table 2. Test length of TIMSS eighth grade mathematics booklets

Booklet	TIMSS 2007	TIMSS 2011	TIMSS 2015
1	29	26	35
2	31	32	33
3	32	32	28
4	29	29	32
5	29	32	34
6	32	33	32
7	33	30	32
8	32	34	30
9	31	34	28
10	32	31	28
11	32	32	29
12	28	32	30
13	28	33	29
14	30	27	30
Mean	30.6	31.2	30.7

Table 2 gives information about the average test length of TIMSS eighth grade mathematics achievement tests, which is about 30 items. The response patterns of Turkey and the USA participants were merged by using anchor items to obtain incomplete data matrix. Data collection design of these assessments is shown in Figure 1.

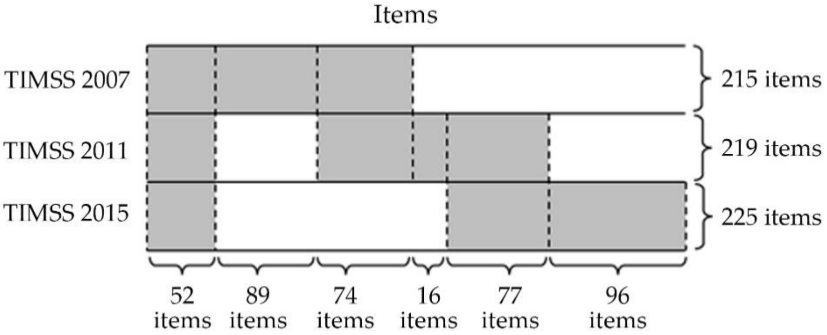

Figure 1. Data collection design of TIMSS eighth grade mathematics assessments

The data matrix contained 45,580 rows (participants) and 404 columns (items). However, 11 of the items (M042273, M062345BA, M062345BB, M062345BC, M062345BD, M062345B, M062342, M062048A, M062048B, M062048C and M062048) were taken out of the analysis since they had all missing responses. Out of the 393 items, dichotomously scored 360 items were calibrated by using 2 Parameters Logistic (2PL) model and polytomously scored 33 items were calibrated by using Partial Credit Model (PCM). In the item pool, all the multiple-choice items were dichotomously scored. However, some of the open-ended items were dichotomously scored and the remaining were polytomously scored. PCM is a unidimensional model for the responses scored in two or more ordered categories (Masters, 2016). MIRT (Glas, 2010) program was used for item analysis and calibrating both dichotomously and polytomously scored items. The item parameter distribution of dichotomously scored items (item difficulty versus item discrimination) is given in Figure 2.

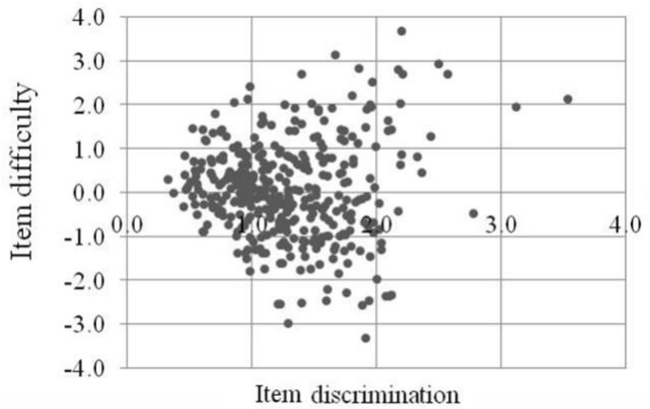

Figure 2. Distribution of item parameters calibrated by 2PL model

In addition to the item parameters, MIRT program also reported ability estimations and standard error values of these estimations based on WML and EAP methods. Statistical information about the ability estimations are given in Table 3.

Table 3. Mean values of ability estimations and standard error values in item calibration

Statistic	Ability estimation method	
	WML	EAP
Ability estimation	-.054	-.063
Mean SE	.371	.328

As shown in Table 3, mean value of ability estimations were -.054 and -.063 for WML and EAP methods, respectively. Also, the mean values of standard errors were .371 in WML and .328 in EAP.

Data Analysis

Test equating and scaling of TIMSS assessments were conducted based on IRT framework (Martin, Mullis & Hooper, 2016) so the assumptions were supposed to be satisfied. The item calibration were conducted based on the unidimensional IRT model by using MIRT software package (Glas, 2010). In this analysis, 360 items were calibrated by 2PL model and 33 items were calibrated by PCM. Afterwards, these item parameters were used in simulation studies. A sample of 1000 simulated test takers were drawn from normal distribution N(0,1) and three sets of simulations were designed. Afterwards, based on the item parameters and drawn ability values, a response matrix having 1000 rows and 393 columns was formed.

In the first set of simulations, variable-length tests were used and .20, .30 and .40 reference values were set for standard error. Next, (a) correlation between true theta and estimated theta, (b) average test length and (c) distribution of item exposure rates, (d) root mean square error (RMSE) and (e) bias were compared for each standard error value. Here, item exposure rate stands for the ratio of the participants facing the item to the total number of participants. For example, if 130 out of 1000 participants saw an item during a test administration, then the item exposure rate for this item would be .13. The RMSE and bias are the values representing the differentiation between predicted (true theta values) and observed (estimated theta values) ability estimations.

Second set of simulations was focused on the comparison of fixed-length tests with 10, 20 and 30 items based on (a) correlation between true theta and estimated theta, (b) mean standard errors and (c) distribution of exposure rates, (d) root mean square error (RMSE) and (e) bias.

Third set of simulations was conducted to indicate the effect of using item exposure control in CAT algorithm whereas the fourth set of simulations were implemented to analyze the efficiency of ability estimation methods.

In these simulations, different number of random items were administered at the beginning of the test as test starting rules, Fisher's information was used as item selection, WML and EAP methods were compared as ability estimation method, variable-length test and fixed-length test were used as test termination rule. Also, the effect of Randomesque method (Kingsbury & Zara, 1989) on the CAT algorithm was examined.

RESULTS

First set of simulations were conducted and 36 conditions were compared to determine the optimum CAT algorithm by comparing three types of starting rules (ability estimations without any constraint i.e. standard version, after three random items or six random items), two different ability estimation methods (EAP or WML) and six different termination rules (fixed-length tests with 10, 20 or 30 items; variable-length tests terminated after reaching .20, .30 or .40 standard error values).

a) Simulations based on variable-length tests

Here, simulations were conducted to compare the effects of the determined situations on variable-length tests so that the average test length and correlation coefficient between true and estimated theta values were calculated. The results are shown in Table 4.

Table 4. Test lengths and correlation coefficients between true and estimated theta in variable-length tests

Method	initial ability estimation	Test termination rule					
		SE < .20		SE < .30		SE < .40	
		test length	r	test length	r	test length	r
EAP	after first item	33	.978	12	.961	6	.931
	after 3 random items	35	.980	13	.955	7	.926
	after 6 random items	36	.978	15	.960	9	.922
WML	after first item	36	.979	13	.953	7	.929
	after 3 random items	36	.980	14	.957	9	.934
	after 6 random items	38	.980	16	.958	10	.933

Table 4 shows that a better measurement precision was obtained with higher correlations but this cost more items as expected. This can be explained by the relationship between the standard errors and the reliability of the test scores. Also, average test length was directly related with the same context. In other words, the algorithm gave more items to the participant so as to reach a standard error less than .20. Decreasing the standard error reference from .40 to .30 almost doubled the test length and tripled when the standard error reference changed from .30 to .20. Using more random items before initial ability estimations increased the test length in variable-length tests. More specifically, variable-length tests needed more items since random items were used in the algorithm rather than selecting the most informative item. Finally, when the effect of EAP and WML ability estimationmethods were analyzed in variable-length tests, there was no prominent differentiation occurs among test lengths and correlation coefficients.

Table 5 indicates the item exposure rate distributions, RMSE and bias of different ability estimations in variable-length tests.

Table 5. Item exposure rate distributions, RMSE and bias of different ability estimations in variable-length tests

Method	Initial ability estimation	Test termination rule	Item exposure rate				RMSE	Bias
			<.01	.01-.20	.21-.40	>.40		
EAP	after first item	SE < .20	0	334	38	21	.210	-.008
		SE < .30	168	205	14	6	.286	-.012
		SE < .40	235	148	6	4	.359	.001
	after 3 random items	SE < .20	0	338	35	20	.209	.000
		SE < .30	163	211	13	6	.296	-.006
		SE < .40	203	181	7	2	.375	.014
	after 6 random items	SE < .20	0	339	34	20	.210	-.004
		SE < .30	150	226	14	3	.280	.017
		SE < .40	214	173	4	2	.382	-.021
WML	after first item	SE < .20	0	333	35	25	.202	-.013
		SE < .30	116	257	15	5	.303	-.017
		SE < .40	198	184	8	3	.383	-.039
	after 3 random items	SE < .20	1	334	35	23	.198	-.019
		SE < .30	138	236	15	4	.284	-.016
		SE < .40	168	215	7	3	.376	-.026
	after 6 random items	SE < .20	1	333	35	24	.202	-.007
		SE < .30	131	245	14	3	.292	-.022
		SE < .40	189	197	5	2	.365	-.015

The effect of variable-length tests and different termination criteria on item exposure rates, RMSE and bias were analyzed and as shown in Table 5, the decrease in the standard error reference value ended up with the decrease in the number of items with underexposure (exposure rates less than .01) This seems to be a positive outcome but at the same time it increased the number of items with overexposure (exposure rates greater than .40). When the effect of variable-length tests on RMSE and bias was examined, stricter test termination rules (smaller standard error reference values) ended up with smaller RMSE and bias.

Although there was no obvious differentiation of EAP and WML methods when RMSE and bias were compared, EAP had less bias. Moreover, WML provided negative bias values in all conditions interpreting that this method had higher observed values (estimated theta) than the predicted values (true theta).

When comparing the test starting rules, it was found that using more random items at the beginning of the test had a positive impact on decreasing the number of items with overexposure (exposure rates greater than .40).

b) Simulations based on fixed-length tests

Second set of simulations were conducted to observe the effect ability estimation methods and starting rules on fixed-length tests containing 10, 20 and 30 items. Table 6 shows the mean standard errors and correlation coefficients between true and estimated theta in fixed-length tests.

Table 6. Mean standard errors and correlation coefficients between true and estimated theta in fixed-length tests

Method	Initial ability estimation	Test termination rule					
		10 items		20 items		30 items	
		mean SE	R	mean SE	r	mean SE	r
EAP	after first item	.310	.946	.231	.973	.194	.980
	after 3 random items	.328	.947	.237	.969	.198	.979
	after 6 random items	.375	.918	.249	.969	.205	.977
WML	after first item	.329	.943	.244	.971	.206	.978
	after 3 random items	.348	.935	.245	.969	.210	.980
	after 6 random items	.430	.912	.260	.966	.212	.976

When Table 6 is examined, the increase in the test length decreased the mean standard error values and increased the correlation coefficients between true and estimated theta. In almost all conditions, an increase in the number of random items at the beginning of a test decreased the correlation coefficients and increased the mean standard errors. In a general perspective, intervening the item selection algorithm has a cost of an increase in test length in order to preserve the reliability. Therefore, using 6 random items at the beginning of the test had smaller correlation coefficients and higher standard error values compared with other two cases (after first item and after 3 random items). Finally, when the ability estimation methods were compared EAP method provided comperatively better results than WML method.

Table 7 shares the item exposure rate distributions, RMSE and bias of different ability estimations in fixed-length tests.

Table 7. Item exposure rate distributions, RMSE and bias of different ability estimations in fixed-length tests

Method	Initial ability estimation	Test termination rule	Item exposure rate				RMSE	Bias
			<.01	.01-.20	.21-.40	>.40		
EAP	after first item	10 items	247	129	11	6	.324	-.020
		20 items	209	142	27	15	.229	.001
		30 items	182	148	36	27	.202	.000
	after 3 random items	10 items	224	154	11	4	.330	-.008
		20 items	193	162	26	12	.246	.006
		30 items	163	171	35	24	.207	.002
	after 6 random items	10 items	233	152	6	2	.398	-.008
		20 items	201	160	24	8	.256	-.012
		30 items	170	170	31	22	.213	-.013
WML	after first item	10 items	236	141	10	6	.334	.005
		20 items	205	146	29	13	.252	.006
		30 items	174	158	34	27	.211	.007
	after 3 random items	10 items	219	160	10	4	.360	-.006
		20 items	189	167	26	11	.246	-.004
		30 items	154	176	37	26	.209	-.001
	after 6 random items	10 items	229	157	4	3	.453	-.007
		20 items	196	164	23	10	.272	-.006
		30 items	165	175	31	22	.223	-.013

In fixed-length tests, longer tests had a positive impact on increasing the number of items having underexposure (exposure rates less than .01) but at the same time had a negative impact on increasing the number of items having overexposure (exposure rates greater than .40). In all cases, RMSE values decreased as the test length increased. Also, administering 6 random items before the initial ability estimation had a positive effect on the item exposure rates.

Although the item exposure rates were different across test lengths with 10, 20 and 30 items, the results were not sufficient to determine the superiority of the ability estimation methods. In other words, EAP and WML methods seemed to have similar item exposure rate distributions. However, when the RMSE values were on focus, EAP provided more comparable results than WML.

Up to this point, fixed-length tests provided better results than variable-length tests. In variable-length tests, especially low and high achievers were given more than 100 items (or even all the 393 items) in order to satisfy termination rule. Even the termination rule could not achieve to decrease the standard error to the set value after implementing all the items in the pool to a participant. On the other hand, the length of the test was 4 items for some of the participants. If an additional minimum and maximum values for test length are not defined in CAT algorithm, using variable-length tests do not seem to be convenient in TIMSS's adaptive testing practices. So, what could be the optimum test length: 10, 20 or 30? In general, an increase in the test length provided evidence to the content validity of the test. Additionally, a test having 10 items did not give stable correlation coefficients across starting rules and ability estimation methods. It seems more reasonable to use tests containing either 20 items or 30 items in CAT algorithm. The scores from a test with 20 items had correlation coefficient of .97 and a test with 30 items test had a correlation coefficient of .98 with true theta. Fixed-length tests containing 20 items had standard error values between .23 and .26 but tests containing 30 items had standard error values between .19 and .21. Hence, tests with 20 items had .93 reliability and tests with 30 items had .96 reliability. The RMSE took values between .229 and .272 in fixed-length tests with 20 items and took values between .202 and .223 in fixed-length tests with 30 items.

When all these results are interpreted, fixed-length tests with 20 items seem to be the optimum condition for CAT algorithm since these tests provide high correlation coefficients and reliability values.

c) Simulations based on item exposure rates

Third set of simulations focus on the item exposure control and the effect of Randomesque method on fixed-length tests having 20 items was analyzed. Based on the results of previous simulations, item exposure rates were defined for each item. These rates were used to decrease the number of overexposed and to increase the number of underexposured items. The results are given in Table 8.

Table 8. Item exposure rate distributions, RMSE and bias of different ability estimations in using item exposure control

Method	Initial ability estimation	Exposure control	Item exposure rate				RMSE	Bias
			< .01	.01 - .20	.21 - .40	> .40		
EAP	after first item	no	209	142	27	15	.229	.001
		yes	197	156	27	13	.245	-.003
	after 3 random items	no	193	162	26	12	.246	.006
		yes	178	175	32	9	.242	.005
	after 6 random items	no	201	160	24	8	.256	-.012
		yes	189	172	25	9	.266	-.008
WML	after first item	no	205	146	29	13	.252	.006
		yes	179	174	31	12	.244	-.006
	after 3 random items	no	189	167	26	11	.246	-.004
		yes	170	189	25	13	.267	-.010
	after 6 random items	no	196	164	23	10	.272	-.006
		yes	183	179	25	11	.274	-.008

According to Table 8, using item exposure control decreased the number of underexposure items (exposure rates less than .01). When RMSE and bias was concerned, there were some differentiation in the values but it did not seem to have a pattern.

Analysis were conducted to determine whether it was more convenient to estimate abilities with either EAP or WML methods and after first item, after 3 random items or after 6 random items. For all of the cases, item exposure rates were calculated for the items in the pool. In order to observe the changes more clearly, these rates were sorted from high to low. The graphs showing the efficiency of item exposure control for different ability estimation methods and starting rules are shared in Figure 3. In the figure, vertical axis stands for the item exposure rates. The horizontal axis indicates the items on which the items with high exposure rates locate to the left and the items with low exposure rates locate to the right.

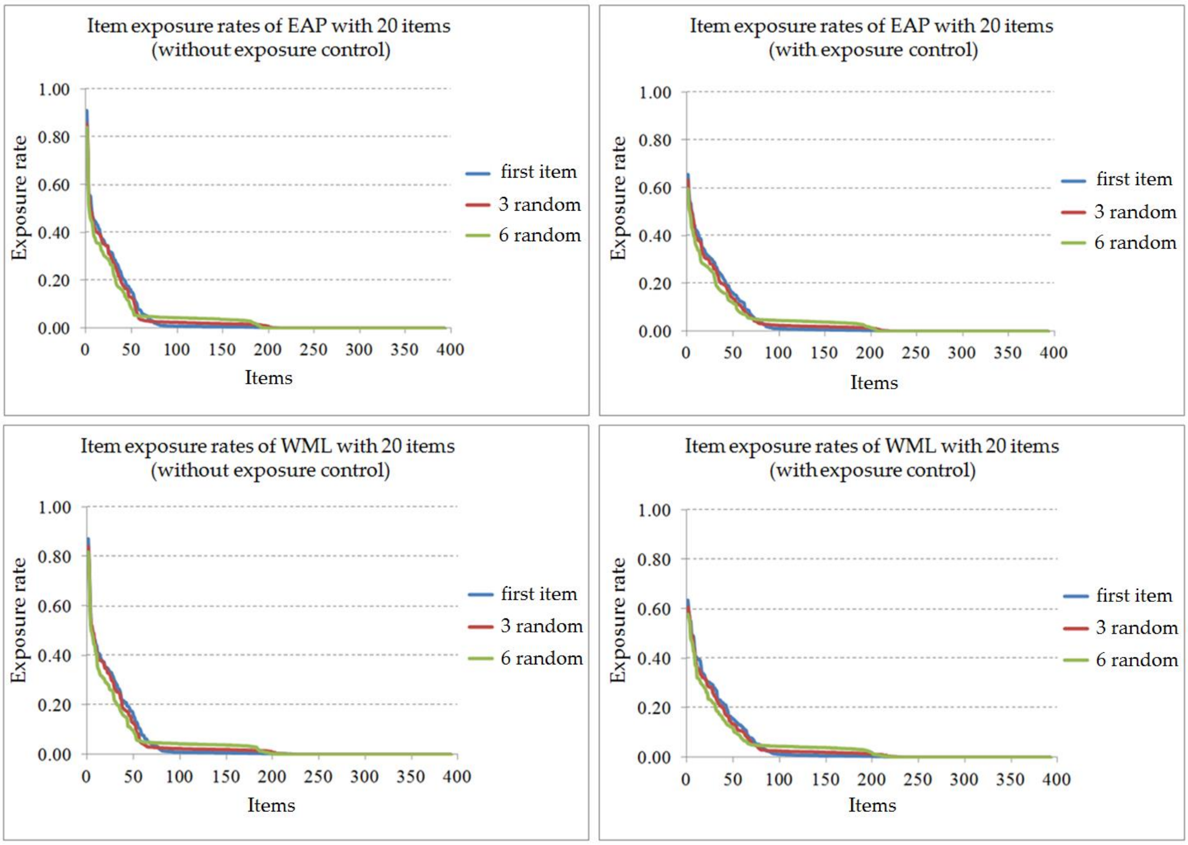

Figure 3. Distribution of item exposure rates by different ability estimation methods either with or without exposure control

In Figure 3, although item exposure control had a positive impact on the item exposure rates of the items in the pool, there was a major problem that almost half of the items were not used in any of the test administrations. When test starting rules were compared, item exposure rates of overexposure items decreased evidently. The main reason behind this is directly related with providing a way to present not used items. Hence, it is believed that using 6 random items at the beginning of the test ensures the effective usage of the item pool so it could be a good starting rule for the optimum algorithm of TIMSS eighth grade mathematics assessments.

Up to this point, the simulation results provide similar results for both EAP and WML.

d) Simulations based on ability estimation methods

Fourth group of simulations were conducted to determine the effectiveness of EAP and WML methods. In the simulations, test starting rule was set to administer 6 random items before initial ability estimation and test termination rule was set to fixed-length tests with 20 items. Moreover, item exposure conrol was used in the comparisons and the relationship between true and estimated theta is given in Figure 4.

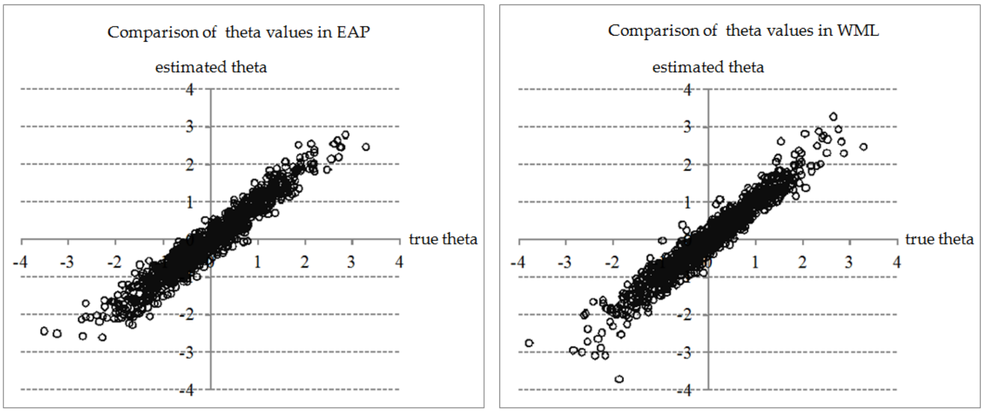

Figure 4. Comparison of estimated theta and true thata in EAP and WML

According to Figure 4, theta values located at very low and very high values in WML were scattered more than as they are in EAP. So, EAP seems to provide a better estimation for the participants from especially quite low and high theta values compared to WML.

To summarize the results of this study, starting ability estimations after six random items as the starting rule, using EAP as the ability estimation method, terminating the test after 20 items and using item exposure control indicated the optimum condition for TIMSS eight grade mathematics assessments. In this case, the mean SE estimation was .253 (.135 as minimum and .468 as maximum) and the correlation between true and estimated theta was .964.

DISCUSSION and CONCLUSION

The aim of this study is to determine the optimum CAT algorithm which is an alternative to the paper and pencil based TIMSS eight grade mathematics assessments. In the simulations, different starting rules, ability estimation methods and termination rules were compared and the effectiveness of item exposure control was analyzed.

As a starting rule, initial ability estimations after first item, after 3 random items and after 6 random items were compared. Although, using more random items at the beginning of the test had a negative effect on RMSE values, its positive impact on the item exposure rates made it indispensible for optimum algorithm. However, it was more convenient to use 6 random items in longer tests. In other words, it was not convenient to use 6 random items in fixed-length tests with 10 items or in variable-length tests with .40 standard error reference because 6 items probably consituted the major part of the test in such cases.

When the ability estimation methods were compared, EAP and WML gave similar results but EAP provided better estimations for especially low and high achievers, which is very similar to the findings of Gu and Reckase (2007).

In order to determine the optimum test termination criteria, variable-length and fixed-length tests were compared. When the standard error was set to .20 in variable-length tests, the correlation coefficients were calculated to be higher but in some of the cases the algorithm presents all the items in the bank but it was not successful to diminish the standard error value below .20. Therefore, it was not practical to use variable-length tests for low achievers and high achievers. To be more specific, then the algorithm could not succeed in decreasing the standard error to .20 even after using all 393 items in the pool. Similar results were interpreted in the study by Gökçe and Berberoğlu (2015). Hence, using a fixed-length test becomes more reasonable in TIMSS eighth grade mathematics assessments. When fixed-length tests with 10, 20 and 30 items were compared, test with 20 items provided more comparable results for TIMSS eight grade mathematics assessments. In the study, Randomesque exposure control was used and the results indicated that this method balanced the item usage by increasing the exposure rates of underexposure items and decreasing the exposure rates of overexposure items. However, in any case, almost half of the items in the pool were not used for any of the participants. In CAT administrations, one of the major problems related with the items is underexposure and overexposure of items (Eggen, 2001; Eggen & Straetmans, 2000). For further studies, it would be better to compare different exposure control methods in TIMSS assesments.

In TIMSS eight grade mathematics assessments, the number of items contained in eighth grade mathematics booklets is about 30. These tests estimated ability with a mean SE value of .328 by EAP method. On the other hand, the optimum CAT algorithm estimated theta values with a mean SE value of .253 with 20 items (with a 35.5% shorter test) by the same method. This results is one of the main advantages of CAT applications. There are many studies indicating that computerized adaptive tests provide more reliable estimations with shorter tests and decreases the testing time (Eggen, 2007; Hambleton et al., 1991; Meijer & Nering, 1999; Mills & Stocking, 1996; Verschoor & Straetmans, 2010).

In all of the cases, there were high correlation coefficients between true that and estimated theta. There are studies reporting that there would be similar ability estimations when different starting rules, ability estimation methods and test termination rules are used in the algorithm (Kalender, 2011; Kezer & Koç, 2014).

This study investigated the applicability of TIMSS eighth grade mathematics assessments as computerized adaptive test and has some limitations. In the literature, starting rules are related with the difficulty of the items at the beginning of the test but instead the effect of starting the test with a group of random items was investigated in this study. Moreover, there are two types of items in the pool either dichotomous or polytomous. In paper and pencil based TIMSS assessments, it is easy to control the number of dichotomous and polytomous items but this study did not focus on balancing item type. In TIMSS eighth grade mathematics assessments, there are 4 learning areas (numbers, geometry, algebra and data-probability) and tests developers can control the number of items for each learning area. However, this study did not consider any control based on content. Finally, open-ended items existed in the item pool of the CAT simulations. Although the ability estimations were carried out by using these items, it would be difficult to use such items in real CAT practices because of their scoring. This is another limitation of the study.

In the study, the data sets of Turkey and United States of America were used. For further studies, the data of other participating countries from TIMSS 1995, 1999, 2003, 2007, 2011 and 2015 mathematics assessments could be analyzed and compared with the results of this study. Also, since this study used eighth grade mathematics data set, further studies could focus on the TIMSS fourth grade mathematics data and check whether to obtain comparable results across grade levels.

REFERENCES

Davey, T., & Pitoniak, M. J. (2006). Designing computerized adaptive tests. *Handbook of Test Development*, 543-574. Routledge.

Eggen, T. J. H. M. (2001). Overexposure and underexposure of items in computerized adaptive testing. *Measurement and Research Department Reports, 1*.

Eggen, T. J. H. M. (2004). Contributions to the Theory and Practice of Computerized Adaptive Testing. Dissertation. Print Partners Ipskamp B.V., Enschede.

Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

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

Glas, C. A. W. (2010) MIRT: Multidimensional Item Response Theory. (Computer Software). University of Twente. Retrieved from https://www.utwente.nl/nl/bms/omd/Medewerkers/medewerkers/glas/#soft-ware

Glas, C. A. W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation*, 35, 83-88.

Gu, L., Reckase M. D. (2007). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control. In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (Vol. 2). Sage.

Kalender, I. (2011). Effects of different computerized adaptive testing strategies on recovery of ability. Yayınlanmamış Doktora Tezi. Middle East Technical University, Ankara.

Kezer, F. & Koç, N. (2014). Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [A comparison of computerized adaptive testing strategies]. *Eğitim Bilimleri Araştırmaları Dergisi - Journal of Educational Sciences Research*, 4 (1), 145-174.

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

Luecht, R. M. & Sireci, S. G. (2012). A review of models for computer-based testing. *Research Report RR-2011-12*. New York: The College Board.

Masters, G. N. (2016). Partial credit model. In *Handbook of Item Response Theory, Volume One* (pp. 137-154). Chapman and Hall/CRC.

Meijer, R. R. & Nering M. L. (1999). Computerized adaptive testing: overview and introduction. *Applied Psychological Measurement*, 23, 187-194.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9 (4), 287-304.

Mullis, I., Martin, V. & Loveless, T. (2016). 20 years of TIMSS, international trends in mathematics and science achievement, curriculum, and instruction. IEA, TIMSS&PIRLS International Study Center Lynch School of Education, Boston College.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., & Hambleton, R. K. (2008). Massachusetts Adult Proficiency Tests Technical Manual, Version 2. *Center for Educational Assessment Research Report No, 677*.

Smits, N., van Straten, A., & Cuijpers, P. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, 188(1), 147-155.

van der Linden, W. J. (1995). Advances in computer applications. In T. Oakland & R. K. Hambleton (Eds.), *International Perspectives on Academic Assessment*, (pp. 105-124). Kluwer Academic Publishers.

van der Linden, W. J. (2001). Computerized test construction. *Research Report*. Twente University, Enschede (Netherlands).

van der Linden, W. J. (2010). Item selection and ability estimation in adaptive testing. *Elements of Adaptive Testing*, 3-30. Springer.

Verschoor, A. J., & Straetmans, G. J. J. (2010). MATHCAT: A flexible testing system in mathematics education for adults. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of Adaptive Testing*, (pp. 137-149). Statistics for Social and Behavioral Sciences. Springer.

Wainer, H. (2000). Computerized Adaptive Testing: A Primer. Mahvah, NJ: Erlbaum.

Zenisky A. L., & Sireci, S. G. (2002) Technological innovations in large-scale assessment, *Applied Measurement in Education*, 15:4, 337-362.