International Journal of Assessment Tools in Education
2019, Vol. 6, No. 4, 555–567

https://dx.doi.org/10.21449/ijate.634091

Published at http://www.ijate.net http://submit.ijate.net/en Research Article

Analyzing the Maximum Likelihood Score Estimation Method with Fences in ca-MST

Melek Gülşah Şahin [1,*], **Nagihan Boztunç Öztürk** [2]

[1] Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey
[2] Hacettepe University, Lifelong Learning Center, Turkey

ARTICLE HISTORY

Received: 30 July 2019
Revised: 01 October 2019
Accepted: 16 October 2019

KEYWORDS

MLEF,
MLE,
EAP,
ca-MST,
Ability estimation

Abstract: New statistical methods are being added to the literature as a result of scientific developments each and every day. This study aims at investigating one of these, Maximum Likelihood Score Estimation with Fences (MLEF) method, in ca-MST. The results obtained from this study will contribute to both national and international literature since there is no such study on the applicability of MLEF method in ca-MST. In line with the aim of this study, 48 conditions (4 module lengths (5-10-15-20) x 2 panel designs (1-3; 1-3-3) x 2 ability distribution (normal-uniform) x 3 ability estimation methods (MLEF-MLE-EAP) were simulated and the data obtained from the simulation were interpreted with correlation, RMSE and AAD as an implication of measurement precision; and with conditional bias calculation in order to show the changes in each ability level. This study is a post-hoc simulation study using the data from TIMSS 2015 at the 8th grade in mathematics. "xxIRT" R package program and MSTGen simulation software tool were used in the study. As a result, it can be said that MLEF, as a new ability estimation method, is superior to MLE method in all conditions. EAP estimation method gives the best results in terms of the measurement precision based on correlation, RMSE and AAD values, whereas the results gained via MLEF estimation method are pretty close to those in EAP estimation method. MLE proves to be less biased in ability estimation, especially in extreme ability levels, when compared to EAP ability estimation method.

1. INTRODUCTION

Individualized tests have been administered together with computer technology for a long time. These tests, also known as Computer Adaptive Tests (CAT), are using Item Response Theory in the background. The relationship between IRT, latent ability and item parameter is continuous and defined with monotonic mathematical function (Embretson & Raise, 2000; Reckase, 2009). In this way, the test administration algorithm is designed so that the test items which are administered to the test taker are adapted in terms of difficulty in line with the test taker's estimated ability while the test is going on. As the individuals receive items appropriate

CONTACT: Melek Gülşah ŞAHİN ✉ mgulsahsahin@gazi.edu.tr Gazi University, Gazi Education Faculty, Department of Educational Sciences, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

to their own level of ability, they do not get the same test form as is the case in pen-and-paper tests. Also, it is prevented that the individuals receive items which are way above or under their estimated ability. A CAT is more effective than a regular test by having the appropriate item pool (Wainer, Kaplan, & Lewis 1992). Using computer technology has provided the users with convenient strategies such as administering and securing the test items as well as analyzing and storing the data easily. Because of the aforementioned virtues, CAT ensures more efficient and precise measurement in individuals' ability distribution. Although CAT has gained a sound ground in terms of application in a variety of fields, it has its own limitations. Some of them can be listed as being difficult to apply in different item formats, requiring a large item pool as well as complicated software and fast computers, not enabling test items to be revised throughout the test, having a complex item selection algorithm, and not being able to get information about psychometric characteristics since test formats are established during the test (Hambleton, Swaminathan, & Rogers, 1991; Hendrickson, 2007; Luecht & Nungester, 1998; Luecht & Sireci, 2011; Sarı, Yahşi Sarı, & Huggins Manley, 2016; Yan, von Davier, & Lewis, 2014).

Due to its limitations, CAT is gradually being replaced by Computer Adaptive Multistage Tests (ca-MST). ca-MST combines the characteristics of linear and adaptive tests. While the appropriate models are chosen according to the individuals' level in the application as is the case in adaptive tests, the test takers can revise the test items as they can do in linear tests and test content is generally set before the test is administered (Leucht & Nungester, 1998). In these tests, there is an adaptation, not on an item basis, but on the basis of item sets called modules (Leucht & Nungester, 1998; Yan, von Davier & Lewis, 2014; Zenisky, Hambleton & Leucht, 2010).

ca-MST is more advantegous than CAT especially because of the fact that the test formats in ca-MST can be examined in advance by test developers and test items can be reviewed by test takers during test administration (Luecht, Brumfield, & Breithaupt, 2006; Hendrickson, 2007).

1.1. Ability Estimation Methods

In individualized tests, which items will be set to a test taker is not decided beforehand. It is necessary to estimate the individual's ability to be able to choose the items. Based on the individuals' performance, the next item which is appropriate to the individual's ability is chosen from the pool which has specific item parameters. Different from CAT, in ca-MST, the individual's ability is estimated after each module and the most appropriate module in the first stage that comes after the estimation is administered to the individual.

Since IRT-based estimation methods are used in estimating the individual's ability, the ones used for CAT can also be used for ca-MST. In the literature, the most frequently used ability estimation methods are Maximum Likelihood Estimation (MLE), Expected a posteriori (EAP), and Maximum a posteriori (MAP) (Baker & Kim, 2004; Embretson & Resie, 2000). MLE method is often chosen because it is based on likelihood function and it provides unbiased estimates. The log likelihood function of an individual which was estimated after an administered test item is represented below.

$$L = \ln(\mu|\theta) = \sum_{j=1}^{n} \left[\mu_j \ln P_j + \left(1 - \mu_j\right) \ln\left(1 - P_j\right)\right]$$

where μ is a response string of j items, which is (μ1, μ2, μ3..), and Pj is the item response function given theta (θ).

In MLE method, the module that provides the most information about the individual is chosen. Although the ML estimator is efficient and unbiased in asymptotical terms, a large item pool is

needed to make use of it while it is not applicable with examinees having all-endorsed or all-not-endorsed response patterns (Embretson & Reise, 2000). Therefore, this method requires individuals to have at least one correct and one incorrect response in order to estimate abilities.

In response to the limitations of MLE, Bayesian-based estimation methods are proposed. These suggested methods include Modal a posteriori – MAP (Samejima, 1969) and Expected a posteriori –EAP (Bock & Mislevy, 1982). The MAP estimator combines the available information in hand and exclusive trait level true for all kinds of possible response patterns. What is problematic about MAP is that it might give biased results when the tests are short (e.g., <20), especially when the prior is used in an incorrect way (Embretson & Reise, 2000).

Contrary to the ML and MAP estimators, EAP estimation of trait level requires a non-iterative process. Unlike ML estimation, EAP yields a finite trait level estimation for all response patterns, including endorsed and not-endorsed ones (Embretson & Reise, 2000). If the item number is finite, the EAP estimator ise biased. The type of bias can be described as that the trait level is biased when the item number is finite (Wainer & Thissen, 1987). In EAP and MAP estimation methods, the item is selected in a way to decrease the individual's ability estimation range to minimum, and ability estimation is done in all kinds of response patterns. Although EAP and MAP estimation methods are similar, there are some significant differences between them. EAP estimation requires a discrete prior contrary to a continuous prior. Because of that, EAP is a scoring strategy that is used most easily among IRT models and testing context (Embretson & Reise, 2000).

In literature, the methods different from MLE and Bayesian-based methods are examined especially for bias reduction (Firth,1993; Magis & Raiche, 2010; Magis, Beland & Raiche, 2010). In summary, each ability estimation method has its own limitations. Han (2016) has developed a method called maximum likelihood estimation with fences (MLEF) to eliminate those method's limitations. Although this method is basically similar to MLE, it requires for score estimation that the MLEF places two imaginary items having fixed responses in order to build ''fences'' around a meaningful range of the log likelihood function. In MLEF, the first imaginary item is accepted to be the lower fence and its b parameter is set at theta, where the lower bound of the theta distribution is expected (e.g., b = -3.5). For the b parameter value, the lower fence should not be higher than any other item included in the test form. Similarly, the second imaginary item is accepted to be the upper fence, and its b parameter is set at u, where the upper log likelihood functions of three-item response patterns bound of the theta distribution is expected (e.g., b = 3.5). The b-parameter upper fence value should be larger than any other item included in the test form. These two ''fence'' items should be established to possess a very high a-parameter value (e.g., a = 3.0). The log likelihood function estimated in MLEF method is presented below.

$$L^* = \ln P_{LF} + \ln(1 - P_{UF}) + \sum_{j=1}^{n} \left[\mu_j \ln P_j + (1 - \mu_j) \ln(1 - P_j)\right]$$

where P_{LF} and P_{UF} are the item response functions of the lower and upper fences.

1.2. Purpose of the Study

Bayesian-based EAP and MAP methods, which are suggested to eliminate some limitations of MLE, one of the most frequently used methods in literature, are known to result in estimates toward the center of a prior distribution, resulting in a shrunken score scale (Weiss and McBride, 1984). There are limited studies about ability estimation methods in ca-MST in literature. One of them is the study carried out by Kim, Moses and Yoo (2015). In that study, researchers have compared MLE, EAP, MAP ve TCF (test characteristic function) methods with different grading methods for tests having different module length. The study is important

as the method, which is previously examined only in CAT, is being examined in ca-MST in different conditions; it provides a comparison of the method with other frequently-used methods in literature; and there is no similar study in the literature. Besides, what makes this study so important is that it compares the method presented by Han (2016) with other existing methods.

The aim of this study is to investigate the effect of MLEF that is developed to eliminate limitations of the related methods on ability estimation in ca-MST. And also, the applicability of MLEF method for ca-MST and the comparison of MLE method that is often used and referred to in literature and Bayesian-based EAP methods for different test conditions are investigated.

2. METHOD

In this study, it is aimed to investigate the effect of different ability estimation methods on ability estimation in ca-MST. For that purpose, real item data were used in the study. Therefore, this study is a descriptive research based on post hoc simulation that uses real item parameters.

2.1. Obtaining of Item Parameters

In the study, an item pool made from TIMSS 2015 mathematics-eight grade assessment items was used. Two item formats are used in the TIMSS assessments: multiple-choice and constructed response. Multiple-choice items represent at least half of the total number of points. Items are grouped into a series of item blocks in TIMSS assessment. Approximately 12–18 items in each block at the eigth grade and a total of 28 blocks were assigned to 14 different achievement booklets at each grade level in TIMSS 2015 (IEA, 2013).

In TIMSS 2015, there are 297 eight grade mathematics items in total. Within the scope of this study, parameters of 159 items in total, which are estimated based on 3-parameter logistics model (Lord, 1980) and graded based on 1-0, are taken from the website https://timssandpirls.bc.edu/timss2015/international-database/. Table 1 shows the descriptive characteristics of parameters belonging to the items handled in the study. Within the scope of the study, an MST item pool is formed with 159 items in total.

Table 1. Mean and standard deviation of item parameters

Parameters	Mean	SD
a	1.31	0.39
b	0.51	0.53
c	0.21	0.08

2.2. Simulee Parameters

In this study, two different distribution types; namely, normal distribution N (0,1) and uniform distribution (-3,3), are examined. Two distribution types are chosen in order to be able to compare MLEF methods with others in case the numbers of simulees, especially at peak ability levels, are different. Therefore, 5000 simulee parameters having both normal and uniform distribution are simulated by using MSTGen (Han, 2013) simulation software tool.

2.3. ca-MST Components

Within the scope of this study, 1-3 (Patsula, 1999; Kim, Moses, & Yoo, 2015) and 1-3-3 (Jodoin, Zenisky, & Hambleton, 2006; Leucht, Brumfield, & Breithaupt; 2006; Park, 2015; Patsula, 1999; Zenisky, 2004) panel designs, which are frequently used in the literature, were studied as in one panel. TIF values used in forming the panels are specified as below in Figure 1.

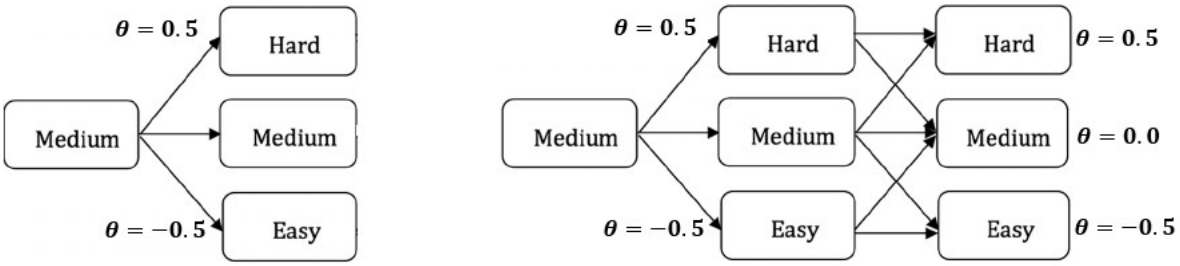

Figure 1. TIF values of 1-3, and 1-3-3 panel designs

In this study, four different module lengths (5-10-15-20) are examined because module lengths vary from small (5 to 10) to large (50 to 100 items) (Luecht, 2000). The lengths of the modules used within the scope of this study are different because test length in ca-MST is correlated with measurement precision (Patsula, 1999), and this study aims to display the effects of ability estimation methods more clearly in long tests.

MLE, EAP and MLEF methods are used for ability estimation of simulees. Maximum Fisher Information is used as item selection method, and "bottom up" is used as test assembly method. For test assembly process, "xxIRT" (Luo, 2017) package program is used in R software (R Development Core Team, 2011).

Table 2 shows 48 conditions examined in this study (2 ability distribution ×2 panel design × 4 module length × 3 ability estimation method).

Table 2. ca-MST components

Components	Variables
Examinee distribution	Normal-Uniform
Panel Design	"1-3"; "1-3-3"
Module Length	5-10-15-20
Estimation method	MLE-EAP-MLEF

2.4. Data Analysis

After MST conditions are created for each variable specified in Table 2, simulee parameters and test conditions are matched up within the context of conditions specified in the study with MSTGen software.

In this study, for each condition, correlation (between the simulated / derived thetas and estimated thetas calculating after ca-MST) root mean square error (RMSE), and average absolute difference (AAD) values are calculated. Pearson's Product Moments Correlation is used in calculating the correlation coefficient. And also, the equations of RMSE and AAD are presented below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}}, \quad AAD = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|}{n}$$

where $\hat{\theta}_i$ represents the estimated level of ability for person i, θ_i represents the known level of ability for person i, and n represents the size of the sample.

In addition to those, it is aimed to examine the changes in ability levels based on bias values in detail. With this aim, ability levels are grouped based on changes of theta 0.5; and bias values are examined in 12 θ change points in uniform distribution and in 15 θ change points in normal distribution (Zenisky, 2004).

3. FINDINGS

In this section, data gathered from the study are presented in two parts. Correlation, RMSE, and AAD values are given in the first part; and conditional bias values are given in the second one.

3.1. Results of Correlation, RMSE and AAD

Correlation, RMSE and AAD values of 48 conditions examined in the study are presented in Table 3. When the correlation values in Table 3 are examined, it is seen that correlation values generally increase when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest correlation value of 0.9679 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the uniform ability distribution having EAP ability level estimation method. On the other hand, the lowest correlation value of 0.6491 is obtained under the condition when the module length is composed of 10 items and the items are administered to simulees in normal ability distribution having MLE ability level estimation method. In panel design 1-3-3, the highest correlation value of 0.9770 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the uniform ability distribution having EAP ability level estimation method. On the other hand, the lowest correlation value of 0.6614 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the normal ability distribution having MLE ability level estimation method.

When ability level estimation methods are examined, it is seen that the highest correlation values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. However, in MLE method, under the condition when module length is composed of 5 items, correlation value is higher than the results under the conditions when module is longer.

It is seen that there is a general increase in correlation values as the number of items in modules increases. Moreover, correlation values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions.

When RMSE values are examined, it is seen that RMSE values generally decrease when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest RMSE value of 6.0165 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest RMSE value of 0.2949 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method. In panel design 1-3-3, the highest RMSE value of 4.2494 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest RMSE value of 0.2515 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method.

When ability estimation methods are examined, it is seen that lower RMSE values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. RMSE values decrease as the number of items in modules increase. Moreover, RMSE values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions. RMSE values obtained via MLEF ability estimation method are found to be a bit higher than the values obtained via EAP estimation method at the two stage panel design as well as the three stage panel design. When compared to MLE, RMSE values obtained via MLEF ability estimation

method can be said to be quite low. Moreover, the lowest RMSE value is obtained at normal distribution at both panel design and all module lengths when MLEF method is adopted.

When AAD values are examined, it is seen that AAD values generally decrease when the panel design shifts from two-stage structure to a three-stage one. In panel design 1-3, the highest AAD value of 4.5277 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest AAD value of 0.2133 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method. In panel design 1-3-3, the highest AAD value of 2.4339 is obtained under the condition when the module length is composed of 5 items and the items are administered to simulees within the uniform ability distribution having MLE ability level estimation method. On the other hand, the lowest AAD value of 0.1812 is obtained under the condition when the module length is composed of 20 items and the items are administered to simulees within the normal ability distribution having EAP ability level estimation method.

When ability estimation methods are examined, it is seen that the lowest AAD values are obtained in EAP ability estimation method. This is valid for both two-stage and three-stage panel designs. It is also valid when the number of items in modules increases. However, the values obtained in both MLEF and EAP ability estimation methods are very close.

AAD values generally increase as the number of items in modules increase. Moreover, AAD values in normal ability distribution conditions are lower compared to the ones in uniform ability distribution conditions.

3.2. Results of Conditional Bias

Conditional bias values calculated in groups according to the changes of 0.5 in ability levels are given in Figure 2. When it is examined, it is seen that under conditions when the test is administered to simulees in normal ability distribution, bias values are higher in extreme ability levels independently of panel design, test length and ability estimation methods. However, the highest bias values in extreme ability levels are under conditions when MLE estimation method is used. While bias values approach the negative infinity as the move is towards -3 ability level, bias values approach the positive infinity as the move is towards +3 ability level. Furthermore, under conditions when two-stage panel design and MLE ability estimation methods are used, more errors are obtained when compared to other methods in mid-levels especially under the condition when the module length is composed of five items. When EAP and MLEF methods are compared independently from stage number, lower bias values are gathered in especially negative extreme points of MLEF method compared to EAP method.

Table 3. Correlation, RMSE and AAD results of ability estimation

PD	AD	AEM	CORRELATION				RMSE				AAD			
			ML5	ML10	ML15	ML20	ML5	ML10	ML15	ML20	ML5	ML10	ML15	ML20
PD1-3	UNIFORM	MLE	0.8517	0.7560	0.7770	0.7655	6.0165	4.3157	4.3010	3.7308	4.5277	2.4887	2.4306	1.9440
PD1-3	UNIFORM	EAP	0.9381	0.9539	0.9645	0.9679	0.7236	0.6226	0.5415	0.5123	0.5453	0.4512	0.3869	0.3623
PD1-3	UNIFORM	MLEF	0.9124	0.9338	0.9501	0.9556	0.7508	0.6377	0.5548	0.5234	0.5421	0.4519	0.3903	0.3654
PD1-3	NORMAL	MLE	0.7518	0.6491	0.6672	0.6664	4.6654	3.3977	3.2046	2.6278	2.6395	1.4965	1.3239	0.9748
PD1-3	NORMAL	EAP	0.8998	0.9321	0.9492	0.9571	0.4438	0.3684	0.3202	0.2949	0.3367	0.2729	0.2356	0.2133
PD1-3	NORMAL	MLEF	0.8621	0.9026	0.9277	0.9381	0.6591	0.5185	0.4345	0.3975	0.4663	0.3510	0.2886	0.2588
PD1-3-3	UNIFORM	MLE	0.7610	0.7591	0.7558	0.7759	4.2494	3.3880	3.2458	2.5211	2.4339	1.7085	1.5556	1.0685
PD1-3-3	UNIFORM	EAP	0.9510	0.9657	0.9709	0.9770	0.6401	0.5327	0.4854	0.4254	0.4711	0.3807	0.3424	0.2966
PD1-3-3	UNIFORM	MLEF	0.9314	0.9516	0.9592	0.9689	0.6507	0.5468	0.5000	0.4365	0.4685	0.3842	0.3464	0.3008
PD1-3-3	NORMAL	MLE	0.6614	0.6832	0.6637	0.7480	3.0762	2.1232	2.2318	1.3407	1.2925	0.7483	0.7713	0.4056
PD1-3-3	NORMAL	EAP	0.9248	0.9515	0.9609	0.9690	0.3871	0.3130	0.2819	0.2515	0.2904	0.2295	0.2038	0.1812
PD1-3-3	NORMAL	MLEF	0.8959	0.9303	0.9420	0.9572	0.5396	0.4242	0.3847	0.3190	0.3741	0.2804	0.2480	0.2087

PD: Panel Desing, AD: Ability Distribution, AEM: Ability Estimation Method, ML: Module Length

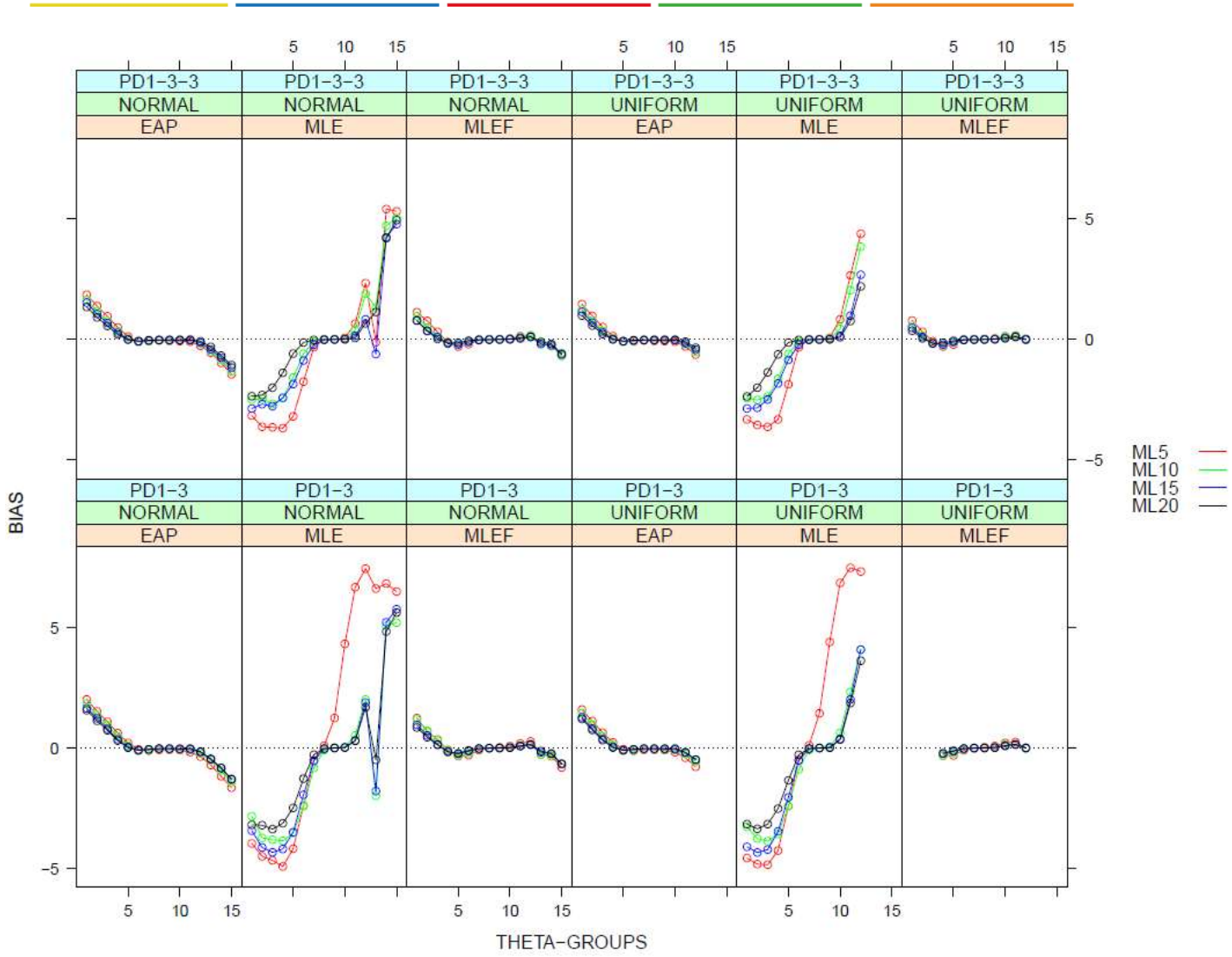

Figure 2. Conditional bias in ability estimation

Similar to the condition when the examinees have normal distribution, under conditions when the items are administered to examinees having uniform ability distribution, the bias values in extreme ability levels are obtained mostly in the one where MLE estimation method is used. Again, the lowest bias value in this condition is obtained with MLEF method.

As the distribution of individuals changes from normal towards uniform, lower bias values are obtained under the condition when the examinees in uniform ability distribution take the test in each ability estimation method. In other words, a good level of estimation can be done under conditions when MLEF method is used.

In general, as the module length changes, the change in bias values can be seen more clearly in MLE method. Especially the bias values obtained in 5-item module length are more noteworthy than the ones in other module lengths. When different module lengths are examined in EAP and MLEF methods, there is a slight change in bias values under conditions when the items are administered to examinees having normal distribution compared to the change in module length especially in extreme points of EAP ability estimation method. However, this situation is less obvious under conditions when MLEF method is used. When the ability distribution is uniform, MLEF ability estimation method has lower bias values in extreme ability levels compared to EAP.

When the graphs obtained from the conditions of different stage numbers are examined, it is seen that there are no significant differences in errors obtained in conditions where EAP and MLEF methods are used. However, lower values are obtained in extreme ability levels of panel design 1-3-3 where MLE method is used. Besides, bias values in mid-ability levels of both uniform and normal ability distributions, where three-stage condition is examined, are in a wider range, around 0.

4. DISCUSSION and CONCLUSION

In this study, the aim is to investigate what results MLEF ability level estimation method, which is brought to literature by Han (2016), gives in terms of ca-MST ability estimation when compared to MLE and EAP methods. In line with this, 48 conditions in total are examined with different panel designs, module lengths and individuals who have different ability level distributions. When the data are interpreted, it is seen that generally MLEF method is more successful in both short and long tests compared to MLE method; and it is successful in decreasing bias values especially in extreme ability levels compared to EAP method.

When correlation, RMSE and AAD values are examined in the study as the indicators of measurement precision, the precision is lower under conditions when MLE ability estimation method is used. Although this result changes as the number of items or stages in modules increases, the results are not close to the values gathered in MLEF or EAP ability estimation methods. It can be said that the result is an expected one considering that there needs to be at least one correct and one incorrect answer in order for MLE estimation method to conduct ability estimation. When the measurement precision values of EAP and MLEF ability estimation methods are examined, it is seen that the values are very close to each other, but EAP method provides a more precise measurement. However, another result is that the differences of correlation, RMSE and AAD values in both methods decrease as the number of items in modules increases. The results are valid for both normal and uniform distribution. When different conditions in two- and three-stage conditions are compared, measurement precision is higher in three-stage conditions. This can be explained by the fact that there is one adaptation point in two-stage tests; in other words, by the fact that there are less measurement results in estimating the simulee's ability level.

When the data gathered in the study are examined in terms of conditional bias values based on ability levels, it is seen that MLE ability estimation method is extremely biased, especially in extreme ability levels. The bias values reach the maximum level, especially in modules where the number of items is five. As the number of items and stages in modules increase, there are slight decreases in bias values. Yen and Fitzpatrick (2006) state in their study that ability estimation whose measurement precision is high can be obtained with MLE method, especially in conditions when the modules are composed of 30 or more items. Besides, it is concluded that it is critical that there are five items in modules for bias values obtained by MLE ability estimation method.

When MLE and EAP ability estimation methods are compared, EAP method shows less bias, especially in extreme ability levels. This result is expected when considering that Bayesian-based methods evolved as a solution to the estimation issues for individuals whose responses are all correct or incorrect in MLE ability estimation method. Similar to the result of this study, Kim, Moses and Yoo (2015) claim that in their two-stage MST study, measurement precision is higher compared to MLE, one of Bayesian-based estimation methods. Also, it is stated that Bayesian-based methods in which MLE is less precise are a better option for high-performing examinees.

When EAP and MLEF ability estimation methods are examined in terms of conditional bias, it is seen that MLEF method has extremely low bias values, especially in extreme ability levels. Therefore, it can be claimed that MLEF method will be slightly biased in estimating abilities, especially of those individuals who have extreme ability levels. Especially when an ability estimation is conducted with ca-MST application for a group whose ability distribution is uniform, bias values are almost around 0 in each ability distribution level. These results are valid even for modules which have the lowest number of items. When Han (2016) compares MLEF method with other estimation methods for CAT in the study, it is stated that similar to this study's results, the estimation can be done with very small bias in extreme ability levels. In line with the results of the study, it is suggested that MLEF method can be preferred over EAP method in terms of providing less biased results. However, especially under conditions when module length is short, it is suggested that test developers can use EAP and MLEF methods for ability estimation instead of MLE method. When deciding on the panel design, since there are more estimation points in three-stage designs, those can be suggested instead of two-stage ones in terms of providing a more measurement precision.

Considering the conditions examined in the study, researchers can further investigate the following issues: trying similar conditions in different panel designs such as 1-2-4; 1-2-2; 1-5-5; 1-2-3-4 where the numbers of items and stages can be changed; examining different ca-MST components such as content balancing or item exposure control; examining similar conditions in different item pools with different item selection methods and examining the effect of different routing module methods on ability level estimation.

ORCID

Melek Gülşah Şahin https://orcid.org/0000-0001-5139-9777
Nagihan Boztunç Öztürk https://orcid.org/0000-0002-2777-5311

5. REFERENCES

Baker, F.B., & Kim, S. (2004). *The basics of item response theory using R.* New York: Marcel Dekker.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Pyschological Measurement*, *6*, 431-444. DOI: 10.1177/014662168200600405

Embretson, S. E., and Reise, S.P. (2000). *Item response theory for pyschologists*. Mahwah, NJ, US: Lawrence Erlbaum

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38.

Magis, D., Beland, S., & Raiche, G. (2010). A test-length correction to the estimation of extreme proficiency levels. *Applied Psychological Measurement, 35*, 91–109.

Magis, D., & Raiche, G. (2010). An iterative maximum a posteriori estimation of proficiency level to detect multiple local likelihood maxima. *Applied Psychological Measurement*, *34*, 75–90.

Han, K. T. (2013). MSTGen: simulated data generator for multistage testing. *Applied Psychological Measurement*, *37*(8) 666–668. doi: 10.1177/0146621613499639

Han, K. T. (2016). Maximum Likelihood Score Estimation Method with Fences for Short Length Tests an Computerized Adaptive Tests. *Applied Psychological Measurement*, *40*(4), 289-301.

Hambleton, R. K., H. Swaminathan and H. J. Rogers. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hendrickson, A. 2007. An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*, 44–52.

International Association for the Evaluation of Educational Achievement (IEA), (2013). *TIMSS 2015 Assessment Frameworks*. Boston College: TIMSS & PIRLS International Study Center, Lynch School of Education.

Jodoin, M. G., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203-220.

Kim, S., Moses, T., & You, H. (2015). A Comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*. 52(1), 70-79.

Luecht, R. M. (2000). *Implementing the Computer-Adaptive Sequantial Testing (CAST) framework to mass produce high quality computer adaptive and mastery tests.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), New Orleans, LA.

Luecht, R.M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229-249.

Leucht, R., & Sireci, S.G. (2011). A review of models for computer-based testing. Research Report. New York: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Luo, X. (2017). *Package 'xxIRT'.* (Version 2.0.3). Retrieved September 25, 2018 from https://cran.r-project.org/web/packages/xxIRT/xxIRT.pdf

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Park, R. (2015). *Investigating the impact of a mixed-format item pool on optimal test design for multistage testing.* (Unpublished doctoral dissertation). University of Texas at Austin.

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Reckase, M.D. (2009). *Multidimensional item response theory*. New York: Springer

Robin, F. (1999, March). *Alternative item selection strategies for improving test security and pool usage in computerized adaptive testing.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montréal, Québec.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monograph Supplement, 34*(4,Pt.2), 100)

Sarı, H.İ., Yahşi Sarı, H., & Huggins Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. DOI: 10.21031/epod.280183

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, *29*(3), 243-251.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Yan, D., von Davier, A.A., & Lewis, C. (2014) *Computerized multistage testing: Theory and applications*. CRC Press

Yen, W. M., & Fitzpatrick, A.R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational measurement.* Westport, CT: American Council on Educaiton and Praeger.

Zenisky, A. L. (2004). Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment. (Unpublished doctoral dissertation). University of Massachusetts at Amherst.

Zenisky, A., Hambleton, R.J., & Luecht, R.M. (2010). Multistage testing: Issues, design and research. In W.J. ven der Linden & C.E.W. Glass (Eds.). *Elements of adaptive testing* (pp.355-372). New York: Springer