Research and Evaluation in Education Journal
Volume 1, Number 1, June 2015 (55-72)
Available online at: http://journal.uny.ac.id/index.php/reid

ESTIMATION OF ABILITY AND ITEM PARAMETERS IN MATHEMATICS TESTING BY USING THE COMBINATION OF 3PLM/GRM AND MCM/GPCM SCORING MODEL

[1)]Abadyo; [2)]Bastari
[1)]Malang State University, Indonesia; [2)]Minsitry of Education and Culture, Indonesia
[1)]aabadyo@gmail.com; [2)]bastari@kemdikbud.go.id

Abstract

The main purpose of the study was to investigate the superiority of scoring by utilizing the combination of MCM/GPCM model in comparison to 3PLM/GRM model within a mixed-item format of Mathematics tests. To achieve the purpose, the impact of two scoring models was investigated based on the test length, the sample size, and the M-C item proportion within the mixed-item format test and the investigation was conducted on the aspects of: (1) estimation of ability and item parameters, (2) optimalization of TIF, (3) standard error rates, and (4) model fitness on the data. The investigation made use of simulated data that were generated based on fixed effects factorial design 2 x 3 x 3 x 3 and 5 replications resulting in 270 data sets. The data were analyzed by means of fixed effect MANOVA on Root Mean Square Error (RMSE) of the ability and RMSE and Root Mean Square Deviation (RMSD) of the item parameters in order to identify the significant main effects at level of $\alpha = .05$; on the other hand, the interaction effects were incorporated into the error term for statistical testing. The −2LL statistics were also used in order to evaluate the model fitness on the data set. The results of the study show that the combination of MCM/GPCM model provide higher accurate estimation than that of 3PLM/GRM model. In addition, the test information given by the combination of MCM/GPCM model is three times higher than that of 3PLM/GRM model although the test information cannot offer a solid conclusion in relation to the sample size and the M-C item proportion on each test length which provides the optimal score of test information. Finally the differences of fit statistics between the two models of scoring determine the position of MCM/GPCM model rather than that of 3PLM/GRM model.

Keywords: *estimation, ability, item parameter, Mathematics test, 3PLM/GRM model, MCM/GPCM model*

Introduction

In 1990s the National Examination in Indonesia was known as *Evaluasi Belajar Tahap Akhir Nasional (EBTANAS)* or, literary translated into English, Final Stage of National Learning Evaluation. The test items for Mathematics in that period were mixed ones consisting of 35 multiple choices and 3 essays. Then, since 1999 such mixed-item format has not been used in the National Examination and, unfortunately, there has not been any proper explanation for such circumstance whereas the use of multiple-choice (M-C) and constructed-response (C-R) test items was heavily implemented in the USA and rhe other countries (Chon, Lee & Ansley, 2007, p.1).

Studies regarding the mixed format of M-C and C-R test items based on the Item Response Theory was popularly conducted in the early 1990 (e.g., Wainer & Thissen, 1993, pp. 103-112 and Lukhele, Thissen & Wainer, 1994, pp. 234-250). These studies then were followed by the other ones conducted by Tang & Eignor, 1997, pp. 1-13; Kennedy & Walstad, 1997, pp. 359-375; Berger, 1998, pp. 248-258; Ercikan et al., 1998, pp. 137-154; Lau & Wang, 1998, pp. 1-13; Garner & Engelhad, pp. 29-51; Li, Lissitz, & Yang, 1999, pp. 1-34; Bastari, 2000, pp. 1-78; Kinsey, 2003, pp. 1-110; Meng, 2007, pp. 1-344; Chon, Lee, & Ansley, 2007, pp. 1-21; Cao, 2008, pp. 1-163; Jurich & Goodman, 2009, p. 3-25; Hagge, 2010, p. 1-284; and He, 2011, pp. 1-174.

Studies regarding the mixed format of M-C and C-R test items mentioned above in general makes use of dichotomous scoring scheme for the M-C test items and of polytomous scoring scheme for the C-R test items. The dichotomous scoring scheme provides two result-possibilities for each item response namely '1' for each correct answer and '0' for each incorrect answer (Bastari, 2000, p. 1; Kinsey, 2003, p. 2; Reynolds, Living-ston, & Willson, 2009, p. 195). On the other hand, the C-R test items are used for gathering information regarding the the incomplete knowledge that perhaps has been possessed by the test participants by demanding the test participants to provide a response toward an item suggestion (for example, the open-ended answers, the short answers and the essays). The C-R test items are usually scored according to the numbers of item completion or the degree of item correctness under the scale of correctness hierarchy (Bastari, 2000, p. 1; Reynolds, Livingston, & Willson, 2009, p. 223). Numerically, the score for the polytomous item depends on the selected IRT model. For example, if the model with K response category is selected then the answer will be scored as 1, 2, ..., K. For each missing data, the MULTILOG provides a category with "0" score.

The dichomotous scoring scheme for the format of M-C test items has several weaknesses because summarizing the incorrect option or the distractors into a certain category might cause the loss of information regarding the score tests. De Ayala (1989, p. 790) states that dichomotization assumed "the test participants act under the principles of knowledge-or-random." As a result, partial knowledge regarding the test participants' trait might be abandoned when the test items are dichomotized and tends to have less accurate in terms of estimation toward the test participants' ability.

There has been an empirical evidence of the selection of distractors in relation to the test participants' characters/traits. The empirical evidence shows that certain distractors might be selected for most of the times by the test participants under different characters/traits (Bock, 1972, p. 29; Levine & Drasgow, 1983, p. 675; Sadler, 1998, pp. 289-290; Thissen, 1976, p. 201; Thissen & Steinberg, 1984, p. 501; Thissen, Steinberg, & Fitzpatrick, 1989, pp. 161-162; Wainer, 1989, p. 192). The evidence support the hypothesis that says that partial information might be attained from the distractors. If the selection of distractors is not related to the test participants' characters/traits, then the opportunity that the test participants have in selecting the distractors will be distributed evenly to all of the available options in all of

the character/trait level based on the principle of equally-likely.

IRT has several models that might the distractors within the multiple choice test items. These models are usually named as the nominal models because, in an a priori manner, these models are not assumed to have sequences among the response items although the relative sequences within the test participants' characters/traits are assumed to exist. The two well-known nominal models for modelling the distractors are Bock NominalModel (Bock, 1972, pp. 29-51) and Thissen Nominal Model (Thissen & Steinberg, 1984, pp. 501-519 – also known as Multiple Choice Model) (DeMars, 2008, p.3). In this study, the researchers would like to review the Multiple Choice Model (MCM) further.

MCM is an expansion of Bock Nominal Model and the Bock Nominal Model is expanded by adding the latent category known as "don't know" or DK (Penfield & Torre, 2008, p.6) which is appropriate for explaining the response-accuracy model in the complex-cognitive tasks (Hoskens& De Boeck, 2001, p. 19). Glasersfeld (1982, p. 613) states that Piaget defined that the cognitive tasks in establishing the knowledge has been related to the outside world and these tasks has been named "cogntive adaptation." One of the domains within the outside world is mathematics which has been developed into the networks of wide abstract hierarchical concepts.

In addition, each individual develops mathematic knowledge within himself or herself through assimilation or accomodation. Since the mind is limited, multiple strategies are used in reducing the mental content including the compression like grouping and naming certain mathematic learning materials. However, the compression gives certain impacts, for example the fraction ¾, division 3 by 4, and the multiplication between ¼ and 3 will be compressed by an individual into a sole object namely ¾. This individual does not consider that the three objects are different.

Such matter might be resolved by means of 'think aloud' method in order to attain the correct answer (Someren, Barnard, & Sandberg, 1994, p. 142; Gierl, Wang, & Zhou, 2008, p. 17). The last matter will be difficult to resolve by using the M-C test items or in general by using the selected-response test item; however, the MCM test items have provided the parameters of guessing proportion in order to accomodate the unexpected aspects.

Kinsey (2003, p. 3) states that there has been a new trend within the recent assessment that has encouraged an increase in the practice of combination among several test items and scoring schemes within a testing format and such trend is known as mixed-format test item. The objective of the combination is to generate a more authentic ability measurement, because the variation of scoring scheme toward the test items might be dichotomous-polytomous or polytomous-dichotomous.

The mixed-format test item for the achievement test often consists of M-C test items and multiple C-R test items (Traub, 1993, p. 30; Wainer & Thissen, 1993, p. 103; Ercikan, et al., 1998, p. 138; Sykes & Yen, 2000, p. 222; Chon, Lee, & Anlsey, 2007, p. 1). The competitive edge of the mixed-item test format or the combination between two test items into a single assessment is that such method might improve both the reliability and the validity or the information of the assessment information (Lau & Wang, 1998, p. 8).

The mixed-format test item consisting of M-C test items and C-R test-items that have been studied by several researchers have not made use of MCM in scoring the M-C test items. For instance, Bastari (2000, pp. 1-78) implemented 3PLM (a dichotomous scoring scheme) for the M-C test items and GRM for the C-R test items within a mixed-format test items. Bastari (2000, p. 54) also recommended the use of 3PLM/GPCM combination for estimating the parameters in the mixed-format test item consisting of M-C and C-R.

Other researchers, such as Kinsey (2003, p. 91) and Chon, Lee, & Ansley

(2007, p. 12), also provided recommendations similar to that of Bastari. The two studies are quite urgent in the assessment development that might generate a more authentic ability measurement and that might improve both the reliability and the validity or the information in the test items and the test formats. In order to achieve this objective, there should be an investigation toward the ability or the performance of the combination among the dichotomous and polytomous IRT model combination in analyzing (especially, in estimating the parameters of) Mathematic mixed-format test items.

A Mathematic test demands the test participants to use mathematic protocols simply for analyzing the problems in the actual world, for designing and determining the resolution strategies and for testing the resolution appropriateness. The test participants should show their understanding toward the mathematic terminologies; in other words, the test participants need the use of definition, algorythm, theorem and other traits for solving a mathematic problem. The test participants are also expected to be able to analyze and interpret the given data (EPAS, 2008, p. 28).

One of the objectives in conducting a Mathematic test is to access the test participants' ability in transferring the qualitative reasoning and the problem-solving skills from one context to another. Therefore, the Mathematic test will continuously be challenged by new situations. The items within the Mathematic test includes four cognitive level namely knowledge and skills, direct application, concept understanding and conceptual integration understanding.

The cognitive development within the Mathematic reasoning and the ability of providing Mathematic evidence are based on the human's basic aspect namely perception, action and language as well as symbolization use that enable us to develop sophisticated and logical options increasingly into the sophisticated knowledge structure. Such matter has been based on what has been called as sensori-motoric language of Mathematics (Tall et al., 2012, p. 1).

Based on the explanation about the cognitive development within the mathematic reasoning, there should be a characterized mathematic test that might be able to capture the pattern of graded response in order to access the mathematic cognitive ability. The polytomous IRT models that have been fit into the patterns of graded response are namely Graded Response Model (GRM), Partial Credit Model (PCM), Generalized Partial Credit Model (GPCM) and Multiple Choice Model (MCM).

The format of convential M-C test items that are generally scored dichotomously make use of 1PLM, 2PLM or 3 PLM; in the Mathematics, the format of conventional M-C test items might also be scored polytomously by using MCM. The underlying paradigm for the perception that the format of conventional M-C test items might also be scored polytomously is that each option is able to describe the gradual partial knowledge up to the option (key) that describes the perfect knowledge or ability. In addition, the MCM is derived from the nominal model. As a result, although the options do not strictly show the gradual partial knowledge, the MCM is still able to perform well within the analysis of M-C items.

The study is an extension toward the parameter estimation of ability and items within the mixed-format test items by considering the recommendations from the previous researchers and by modifying the scoring scheme. Then, the scoring modification is emphasized on the M-C test item format, which previously makes use of 3PLM and then makes use of MCM. The change of the scoring scheme is still linear; in mathematical terms, the researchers would like to show that the 3PLM is one of the MCM derivations.

A study which was conducted by Bastari (2000, pp. 1-78) made use of a 3PLM/GRM combination in order to estimate the relationship in the mixed-item test format in the common scale. On the

other hand, the study makes use of a MCM/GPCM combination in order to estame the parameters of ability and items in the mixed-item mathematic test format. Due to the change of the scoring scheme, the main problem that will be discussed in the study is 'How is the performance of MCM/GPCM combination in comparison to that of 3PLM/GRM combination in analyzing the mixed-item mathematic test format?'

In order to attain the answers toward the main problem of the research, the researchers conducted a study regarding the influence of 3PLM/GRM combination and that of MCM/GPCM toward: (1) the accuracy in the ability estimation parameters and in the test item estimation parameters; (2) the optimalization of Test Information Function (TIF); (3) the derivation of estimation standard errors; and (4) the compability of the combination between the two models into the data (the data fitness) for the various proportion of M-C and essay items, the test length and the sample size. The scoring model combination, the M-C and essay test items proportion, the test length and the sample size are the factors that will be manipulated in the study. Finally, the results on the performance of the combination between the two models will be compared in order to find which combination that is superior to another.

Method

The study was a simulation one by implementing the fixed effect factorial design 3 x 3 x 3 x 2. The first factor consisted of three types of M-C and essay test tems (75:25, 80:20, and 90:10). The second factor consisted of three types of test length (which has been considered in the context of sub-summative test, the National Examination and the aptitude test namely 20 items, 40 items and 60 items respectively). The third factor consisted of three sizes of sample simulation (400, 1000, 3000). The fourth or the final factor consisted of two combinations of scoring scheme namely the 3PLM/GRM combination and the MCM/GPCM combination.

The study was conducted in the Educational Research and Evaluation Laboratory, the Computer Laboratory and the Graduate Program Library of Yogyakarta State University. The study was conducted for almost one year, starting from August 2013 until June 2014.

Within August until December 2013, a syntax for PARSCALE and MULTILOG software was developed. In relation to the syntax development, there had been a use of standardized normally distributed θ ability data with 1000 test participants by means of WinGen2. Based on the θ ability data that had been assumed as the true ability (true theta), the researchers found responses toward 20, 40 and 60 test items for about 54 x 5 = 270 data assembly according to the design of data attainment and the data assembly was replicated by means of WinGen2 as well. However, the data of the answer responses might not be run in the PARSCALE software.

Finally, on January 2014 the researchers decided to attain the simulation data by running the MS Excel 2007 software based on the response data of 2003 Junior High School Examination for the Mathematics in the Province of Yogyakarta Special Region. The data attainment was performed by the researchers themselves with the following phases.

First, the researchers performed a unidimensional assumption test by using exploratory factor analysis (EFA) toward the Mathematic test items of 2003 Junior High School National Examination. At the beginning, the 40 test items of the National Examination did not meet the unidimensional assumptions. After the data had been reduced repetitively, the researchers found 33 items that met the unidimensional assumptions and these items were shown in the following Scree Plot within the Figure 1.

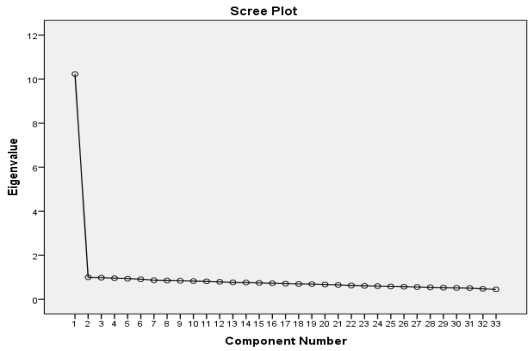

Figure 1. Scree Plot of EFA Results toward 33 Test Items of 2003 Junior High School Mathematic National Examination

Second, based on the data of the responses toward the unidimensional test items found by running the MS Excel 2007 software, the researchers found the θ ability by running the PARSCALE 4.1 software and the θ ability was assumed as the true theta. The θ distribution normality test was performed by running the MINITAB 16 software and the results of the test showed that the θ distribution was not normal. After the researchers performed the data editing process, the researchers found that there had been many outliers that caused the distribution to be asymmetrical. These outliers were shown by the asterisks in the boxplot within the Figure 2.

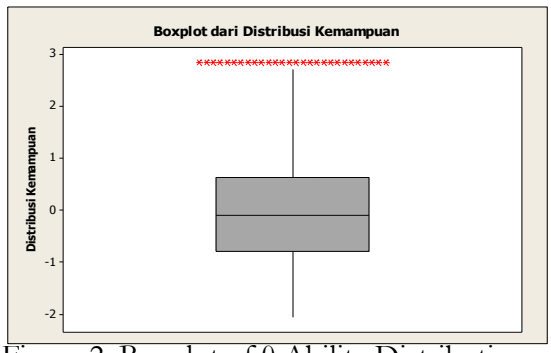

Figure 2. Boxplot of θ Ability Distribution from the Test Participants of 2003 Junior High School Mathematics National Examination

By reducing several scores, including the extreme ones, and then by performing the distribution normality test repetitively, eventually the researchers found a normal distribution with the mean 0.1466 and the standard deviation 0.8803 from the ability of the 2323 test participants at that year. These findings were shown by the results of Anderson-Darling normality test in Figure 3.

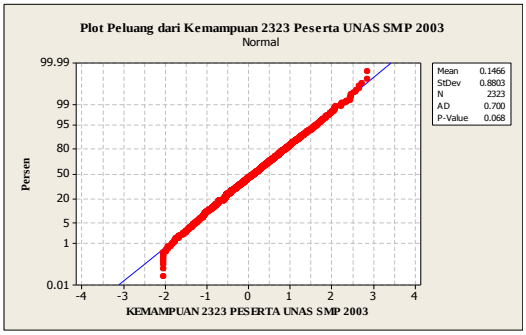

Figure 3. Anderson-Darling Normality Test

Third, the size of the random samples, namely 400, 1000 and 3000 was taken from the θ ability normal distribution shown in Figure 3 by implementing random sampling techniques with replacament and by running the MINITAB 16 software. These samples contained response data found through the operation of MS Excel 2007 software in the second phase. The data was the results of scoring through the 3PLM/GRM combination and the MCM/GPCM combination in terms of test length, the M-C test items proportion and the essay which variations had been mentioned previously.

In addition to the phase of initiating the response toward MCM, the researchers performed checking toward the order of M-C item option by employing MULTILOG software. The order was based on the score of relative frequency or the opportunity of answering the option that a high-level individual had. These scores were described by the Item Characteristic Curve (ICC) of the test items.

For example, test item number 12 had four options. The MULTILOG software employed the code '0' for the missing data and '1' for the 'don't know' or DK response; therefore, the response code for the answer 1, 2, 3, 4 was shifted into 2, 3, 4, 5. Figure 4 showed the ICC of test item number 12. Curver 5 described the opportunity that the test participants had in responding the highest category (the answer key). The rest of the curves, namely the curver 4, 3, 2 described the opportunity that the test participants had in responding the distractors' category which level of truth was

below the correct answer. Paying attention to the high-level ability or the above-average ability, the order of opportunity score (described by the order of the curve) had been clear. If the test item had not been good, then the order would be difficult to determine or even would not be found.

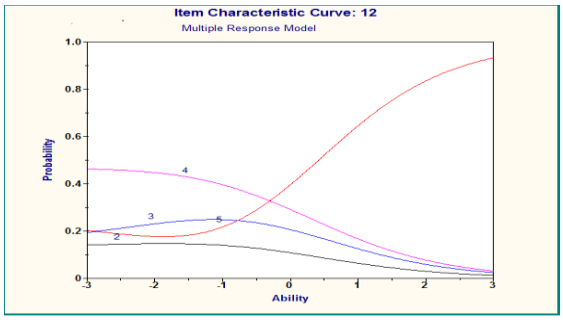

Figure 4. ICC of Test Item Number 12

The control variables within the simulation study were the scoring model combination, the M-C test and essay test items proportion, the simulation sample size, and the test length. Then, the response variables were the test item parameter ability accuracy, the TIF and the estimation standard error (S.E (θ) and S.E (PAR).

The response data gathered from 270 data assembly that had been found was given PRN extension. Each data assembly was run in the PARSCALE 4.1 software by using the syntax that had been developed previously. The outputs generated from the calculation by running the PARSCALE 4.1 software for each design combinaton were the ability estimates (theta estimates), the item parameter estimates (slope estimates, location estimates and guessing parameter estimates) and 2LL statistics. The parameter estimation accuracy was evaluated by using the criteria of root mean squared error (RMSE) and the root mean square differences (RMSD) methods.

Findings and Discussions

The study was conducted to answer four research questions, namely how is the influence of 3PLM/GRM combination and the MCM/GPCM combination toward: (1) the accuracy in the ability estimation parameters and in the test item estimation parameters; (2) the optimalization of Test Information Function (TIF); (3) the derivation of estimation standard errors; and (4) the compability of the combination between the two models into the data (the data fitness) for various proportion of M-C and essay items, the test length and the sample size.

The fixed effect MANOVA was run in the RMSE (θ), the RMSE (PAR), the RMSD (slope), the RMSD (location) and the RMSD (guessing) upon the main effects of the model, sample, proportion and test length in order to answer the first question. The researchers only investigated the significance of the main effect because the interactions of the main effect were incorporated into the statistical testing error since each cell from all factor combinations only contained one datum (Bastari, 2000, p. 31). The effect size from these significant factors were evaluated by using the value of partial eta square (η^2) and the Cohen's criteria (1988) which states that if the score of η^2 = 0.1; 0.25; 0.4, the factor influence respectively will be small, moderate and big. The MANOVA in the study employed the significance level α = 0.05.

The results of MANOVA show that the Pillai's Trace and the Wilks' Lambda statistical scores are significant except for the test length and these results are presented in the Table 1. The results of MANOVA show the p-values for the main effects with RMSE as the dependent variable for (θ) and (PAR). It has been apparent in the Table 1 that all of the main effects, except the test length, have significant F score. On the other hand, for RMSE (θ) the scoring model factor, sample size, M-C/C-R test item proportion has η^2 scores respectively as follows: 0.213; 0.480; 0.196. These scores imply that the sample size is the only factor that has big influence while the scoring model and the M-C/C-R test item proportion respectively has moderate and small influence. For RMSE (PAR) the η^2 scores are respectively as follows: 0.474; 0.730; 0.268. Therefore, the test item proportion is the only factor that has moderate influence while the sample size and the scoring model are the factors that have big influence.

Table 1. *p* values from the Results of MANOVE for RMSE

Source	df	(θ)	(PAR)
Scoring Model	1	**0.001**	**0.000**
Sample Size	2	**0.000**	**0.000**
M-C/C-R Item Proportion	2	**0.007**	**0.001**
Test Length	2	**0.095**	**0.774**

Note:
df = degree of freedom
p-values printed in bold meant that the F values are significant at the level $\alpha = 0.05$

In order to ease the interpretation toward the results of MANOVA, the researchers performed a graphic analysis from the plots that state the comparison between the results of RMSE (θ) marginal mean estimates and those of RMSE (PAR) marginal mean estimates in terms of the 3PLM/GRM scoring model and the MCM/GPCM scoring model according to the sample size, the M-C/C-R item proportion and the test length. Figure 5, 6 and 7 depicted the results of RMSE (θ).

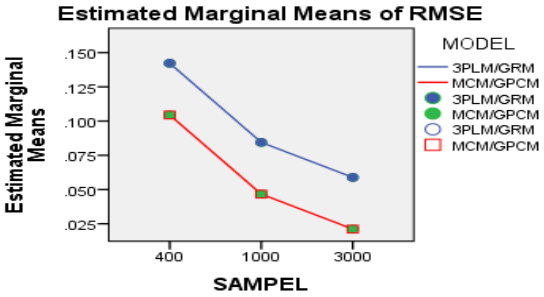

Figure 5. RMSE (θ) Marginal Mean Estimates according to the Sample Size

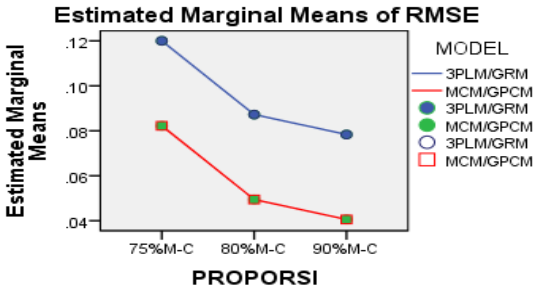

Figure 6. RMSE (PAR) Marginal Mean Estimates according to the M-C/C-R Test Item Proportion

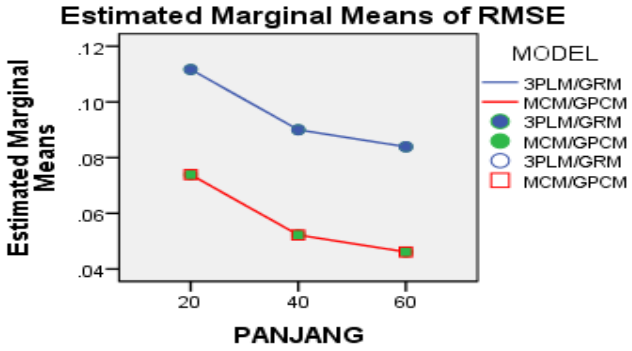

Figure 7. RMSE (θ) Marginal Mean Estimates according to the Test Length

Seen from Figure 5, 6 and 7, it has been apparent that the scores of RMSE (θ) marginal mean estimates in terms of MCM/GPCM scoring model are smaller than those of 3PLM/GRM scoring model. The finding gives implication that the combination of MCM/GPCM provides high accuracy in estimating the RMSE (θ) marginal mean estimates than that of 3PLM/GRM. Furthermore, it has also been apparent that the bigger the sample size and the M-C test item proportion and the longer the test length are, the smaller the scores of RMSE (θ) marginal mean estimates would be. The finding implies that the bigger the sample size is, the M-C test items would be in the mixed-format test item and the longer the test length is, the more accurate the mixed-format test item would be in estimating the RMSE (θ) marginal mean estimates. Meanwhile, Figure 8, 9 and 10 depict the results of RMSE (PAR).

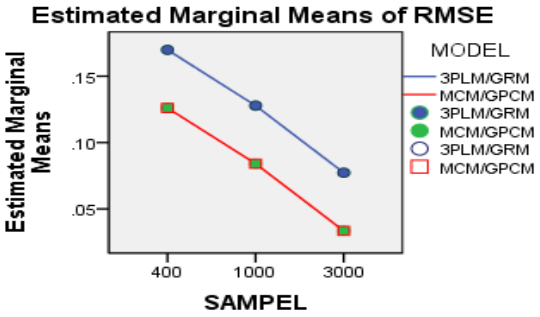

Figure 8. RMSE (PAR) Marginal Mean Estimates according to the Sample Size

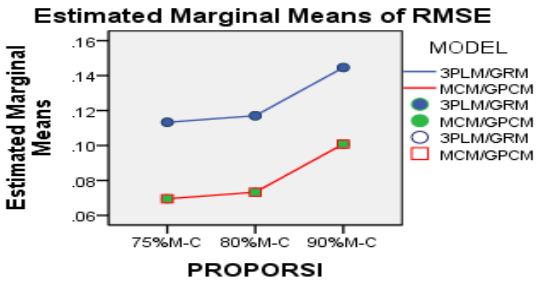

Figure 9. RMSE (PAR) Marginal Mean Estimates according to the M-C/C-R Test Item Proportion

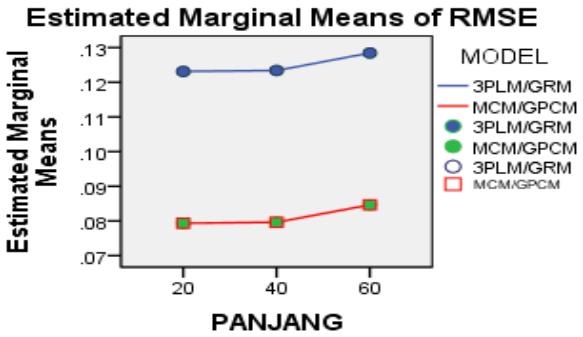

Figure 10. RMSE (PAR) Marginal Mean Estimates according to the Test Length

Similar to Figure 5, 6 and 7, Figure 8, 9 and 10 it has been apparent that the scores of RMSE (PAR) marginal mean estimates in terms of MCM/GPCM combination are smaller than those of 3PLM/GRM combination. The finding implies that the MCM/GPCM combination provides higher accuracy in estimating the RMSE (PAR) marginal mean estimates than the 3PLM/GRM combination. However, for the M-C test item proportion and the test length, the estimation accuracy is reversed, namely, the smaller the sample size is, the more accurate the result would be.

The results of MANOVA for RMSD show that the Pillai's Trace and the Wilks' Lambda statistic score are significant, except for the test item proportion. These statistic scores are presented in Table 2 and show the *p*-values for the main effects with RMSD as the dependent variable for the slope, the location and the guessing. In Table 2, it is clear that for the RMSD (slope), all of the main effects have significant F value. The η^2 values for the model factor, the sample size factor, the proportion factor and the test length factor, respectively, are 0.368; 0.536; 0.167, 0.224. The RMSD (location) is similar to the RMSD (slope) and the η^2 values for the RMSD (slope) respectively are 0.208; 0.604; 0.147; 0.382. Finally, for the RMSD (guessing) the M-C test item proportion and essay test item proportion are the only factors which F values are not significant and the only factor that have big influence is the sample size with the η^2 values = 0.604.

Table 2. *p*-values from the results of MANOVA for RMSD

Source	df	Slope	Location	Guessing
Scoring Model	1	**0.000**	**0.001**	**0.028**
Sample Size	2	**0.000**	**0.000**	**0.000**
M-C/C-R Item Proportion	2	**0.015**	**0.026**	**0.142**
Test Length	2	**0.003**	**0.000**	**0.000**

Note:
df = degree of freedom
p-values printed in bold meant that the F values are significant at the level $\alpha = 0.05$

The graphic analysis toward the plots shows the comparison between the results of RMSD (slope), RMSD (location) and RMSD (guessing) marginal mean estimates in terms of 3PLM/GRM scoring model and of MCM/GPCM scoring model according to the sample size, the M-C and essay test item proportion and the test length. The graphic analysis was conducted in order to ease the interpretation toward the results of MANOVA. Figure 11, 12 and 13 depict the results for RMSD (slope).

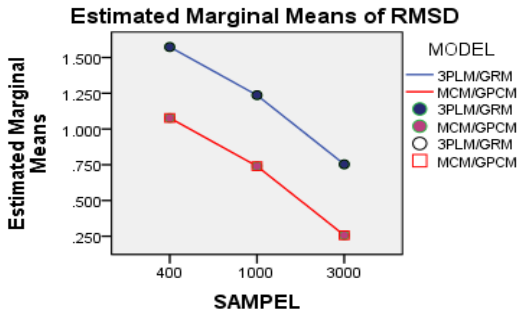

Figure 11. RSMD (slope) Marginal Mean Estimates according to the Sample Size

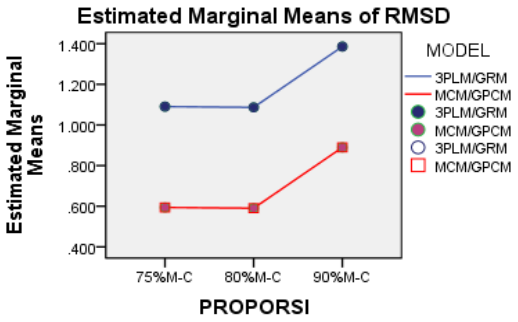

Figure 12. RSMD (slope) Marginal Mean Estimates according to the M-C/C-R Item Proportion

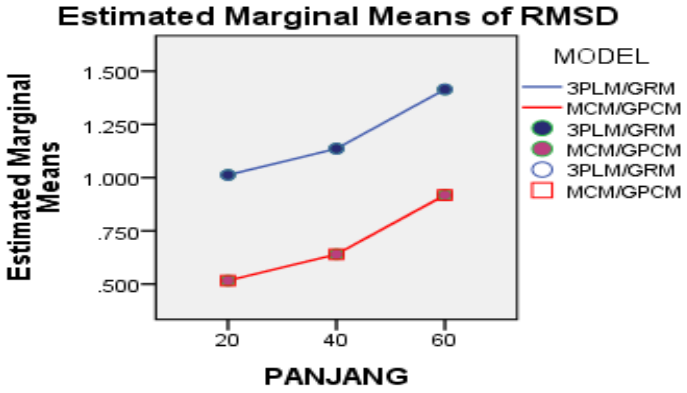

Figure 13. RSMD (slope) Marginal Mean Estimates according to the Test Length

The graphic analysis toward the plots shows that the comparison between the results of RMSD (location) and those of RMSD (guessing) is similar to the graphic analysis toward the RMSD (slope). As a result, overall, the graphic analysis from the RMSE (θ) until the RMSD (guessing) shows that both the RMSE and the RMSD marginal mean estimates that have been analyzed by means of MCM/GPCM combination are smaller than those of 3PLM/GRM combination. The finding implies that the MCM/GPCM combination is more accurate in estimating both the θ ability parameter and the test item parameter.

The second research question was related to the optimalization of Test Information Function (TIF). In order to find the optimal values of Test Information Function from each test length upon the various M-C and essay test item proportion and the various sample size, the researchers drafted the list of the optimal values in Table 3. Table 3 also contains the θ value range in which the maximum score of TIF would be found. Finally, the researchers made a comparison between the optimal values derived from the TIF and the optimal values derived from the 3PLM/GRM method toward the MCM/GPCM method for the test length 20, 40 and 60 and the scores are respectively as follows: 0.364583; 0.358974;

and 0.348485. The values in the comparison show that the TIF optimal values given by 3PLM/GRM method are almost one-third from those of MCM/GPCM method.

Table 3. The Comparison of Optimal Values in the Total Test Information from the Combination of 3PLM/GRM Model and the Combination of MCM/GPCM Model

TEST LENGTH	SAMPLE SIZE	PROPORTION	RANGE (θ)	MODEK	
				3PLM/GRM	MCM/GPCM
20	400	75%	- 0.4 to – 0.3	13.5	30.5
20	400	80%	- 0.4 to – 0.3	17.5	34.5
20	400	90%	- 0.4 to – 0.3	15.5	48.0
20	1000	75%	- 0.4 to – 0.2	16.0	32.0
20	1000	80%	- 0.4 to – 0.2	16.0	36.0
20	1000	90%	- 0.4 to – 0.2	12.0	38.0
20	3000	75%	- 0.4 to – 0.2	13.5	32.5
20	3000	80%	- 0.4 to – 0.2	14.5	33.0
20	3000	90%	- 0.4 to – 0.2	13.5	36.0
40	400	75%	- 0.6 to – 0.3	28.0	69.0
40	400	80%	- 0.6 to – 0.3	27.5	76.0
40	400	90%	- 0.6 to – 0.3	26.5	75.0
40	1000	75%	- 0.4 to – 0.3	25.0	78.0
40	1000	80%	- 0.4 to – 0.3	26.0	74.0
40	1000	90%	- 0.4 to – 0.3	26.5	72.0
40	3000	75%	- 0.4 to – 0.3	26.0	78.0
40	3000	80%	- 0.4 to – 0.3	25.0	68.0
40	3000	90%	- 0.4 to – 0.3	26.0	72.0
60	400	75%	- 0.4 to – 0.3	34.0	110.0
60	400	80%	- 0.4 to – 0.3	42.0	125.0
60	400	90%	- 0.4 to – 0.3	46.0	124.0
60	1000	75%	- 0.4 to – 0.2	38.0	116.0
60	1000	80%	- 0.4 to – 0.2	36.0	105.0
60	1000	90%	- 0.4 to – 0.2	46.0	132.0
60	3000	75%	- 0.4 to – 0.3	39.5	118.0
60	3000	80%	- 0.4 to – 0.3	38.0	116.0
60	3000	90%	- 0.4 to – 0.3	38.0	119.5

These values are presented visually in the line graphic of Figure 14. The symbols which are used in the legend of Figure 14 resemble the following meaning. The numbers in the square brackets show the test length. The symbol before the square bracket resembles the model combination which was employed by the researchers in the scoring scheme according to the sample variation and the M-C and essay test items. For instance, MCM/GPCM[60] refers to the results of 60-item test length that were scored by means of MCM/GPCM combination for the sample size of 400, 1000 and 3000 and the M-C test item proportion is 75%, 80% and 90%. Based on the scores of the ratio above and from the results which are depicted in Figure 13, it has been apparent that the optimal values of TIF that had been analyzed by means of MCM/GPCM combination are three times higher than those of 3PLM/RM combination.

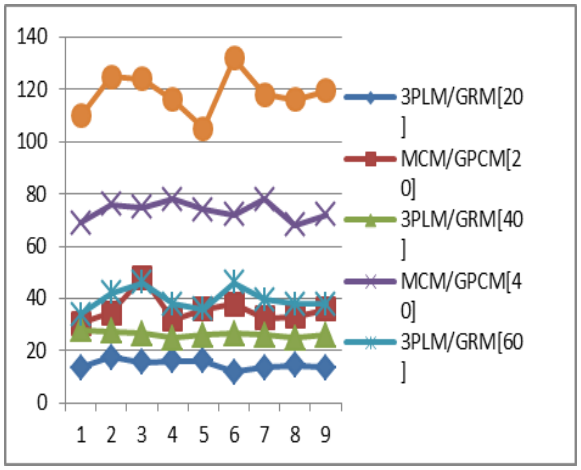

Figure 14. The TIF Optimal Values according to the Test Length and the Sample Size as well as the M-C and Essay Test Item Proportion

The third research problem was related to the estimates' standard error derivation. Similar to the first research problem, in order to answer the third research problem, the researchers performed fixed effect MANOVA on the RMSE-S.E (θ) and the RMSE-S.E(PAR). Table 4 contains the p-values of MANOVA on the RMSE-S.E. for (θ) and (PAR).

Table 4. *p*-values from the Results of MANOVA for RMSE-S.E

Source	df	(θ)	(PAR)
Scoring Model	1	**0.027**	**0.000**
Sample Size	2	**0.000**	**0.000**
M-C/C-R Item Proportion	2	**0.004**	**0.016**
Test Length	2	**0.558**	**0.715**

Note:
df = degree of freedom
p-values printed in bold meant that the F values are significant at the level $\alpha = 0.05$

The results of MANOVA in the RMSE.S-E are similar to those of MANOVA in the RMSE for (θ) and (PAR) with the significant Pillai's Trace and Wilks' Lambda values, except for the test length. It has been apparent in Table 4 that all of the main effects, except the test length, have significant F values. Meanwhile, for the RMSE-S.E (θ) the scoring model, the sample size, the M-C and essay test-item proportion have η^2 values respectively as follows: 0.102; 0.530; 0.217. These values imply that the sample size is the only factor that has big influence while the scoring model and the M-C and essay test item are the factors that have small influence. For the RMSE-S-E (PAR) η^2 values respectively as follows: 0.340; 0.517; 0.164. Therefore, the sample size is the only factor that has big infuence while the scoring model and the M-C and essay test item proportion are the factors that have moderate and small influence.

The graphic analysis toward the plots state the comparison between the RMSE-S.E (θ) marginal and the RMSE.S-E (PAR) mean estimates results by means of 3PLM/GRM and MCM/GPCM according to the sample size, M-C and essay test item proportion and the test length. These values are depicted in Figure 15, Figure 16, Figure 17, Figure 18, Figure 19 and Figure 20.

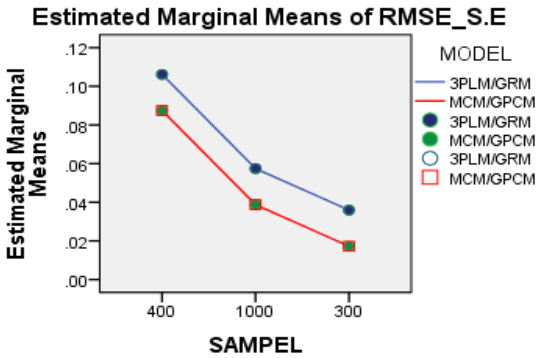

Figure 15. RMSE.S-E (θ) Marginal Mean Estimates according to the Sample Size

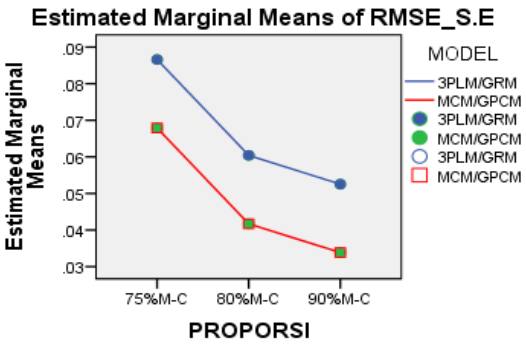

Figure 16. RMSE.S-E (θ) Marginal Mean Estimates according to the M-C and Essay Test Item Proportion

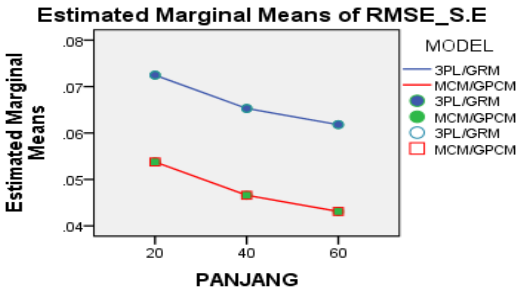

Figure 17. RMSE.S-E (θ) Marginal Mean Estimates according to the Sample Size

The results of graphic analysis for RMSE.S-E (θ) are similar to those of RMSE (θ) and there had been consistency that the bigger the sample size is, the bigger the proportion and the longer the test length the more accurate the estimates would be. Similarly, for the graphic analysis of RMSE.S-E (PAR), there has been consistency with the graphic analysis of RMSE (PAR), namely the smaller the M-C test item proportion and the shorter the test length the more accurate the estimates would be toward the RMSE-S.E marginal mean estimates.

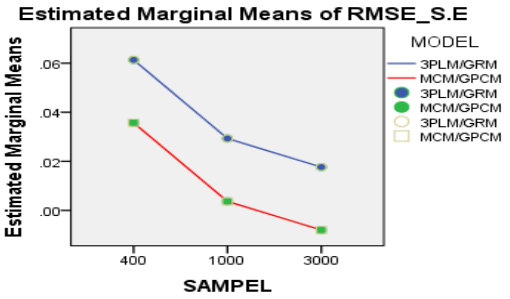

Figure 18. RMSE.S-E (PAR) Marginal Mean Estimates according to the Sample Size

The fourth problem formulation was related to the compability of both combinations to the data fit. In order to evaluate the model or the combination compatibility to the data fit, the researchers made use of minus 2 log likelihood (–2LL) statistic which had chi-square (χ^2) distribution. The big values from the -2LL statistic show that the model has been less compatible to the data. In order to compare which model might be compatible to the data fit, the researchers made use of the gap between the two -2LL statistics which also has chi-square (χ^2) distribution. The values which were generated from the gap between the -2LL (3PLM/GRM) statistic and the -2LL (MCM/GPCM) statistic are presented in the column 'GAP -2LL(χ^2)' in Table 5.

All of the *p*-values in the colum 'p-VALUE' of Table 4 are not equal or bigger than 0.05 and all of the -2LL values for the 3PLM/GRM model are bigger than those of MCM/GPCM model; therefore, it can be concluded that the MCM/GPCM model is more fit to the data than the 3PLM/GRM model. Therefore, from the four types of data analysis employed for solving the four problem formulations, it has been apparent that the performance of MCM/GPCM combination is more superior than that of 3PLM/GRM in analyzing the Mathematics mixed-item test format.

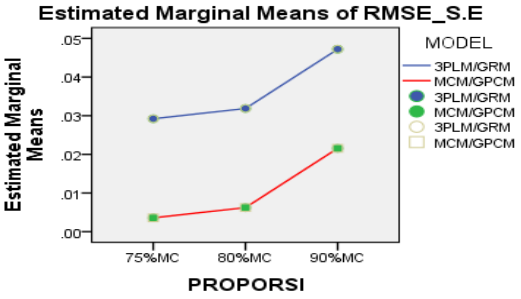

Figure 19. RMSE-S.E (PAR) Mean Marginal Estimates according to the M-C/C-R Item Proportion

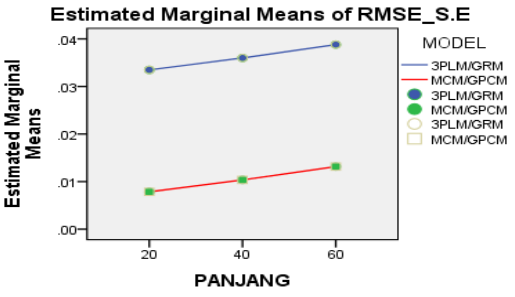

Figure 20. RMSE-S.E (PAR) Mean Marginal Estimates according to the Test Length

Actually, the scoring by MCM/GPCM model in overall implemented many categories and within the study both MCM and GPCM made use of the four categories or both of the models made use of polytomous score. On the other hand, the scoring by 3PLM/GRM model made use of mixed categories namely two categories were mixed with multiple categories (in this study four categories). As a result, the bigger the M-C item proportion in the 3PLM/GRM scoring model is, the more items would be scored by means of two categories or by means of dichotomous score. The finding supports the one from the previous studies (Wasis, 2009, p. 104; Kinsey, 2003, p. 87; Si, 2002, p. 77) which state that the scoring under the polytomous manner will generate better estimates for the test participants' ability than the dichomotomous manner.

Table 5. Comparison between the 3PLM/GRM Model and the MCM/GPCM Model in Terms of -2LL Statistics.

Test Length	Sample Size	Proportion	Model		−2LL (χ^2) GAP	df	P-values	Better Fit
			3PLM/GRM −2LL	MCM/GPCM −2LL				
20	400	75%	16026.2	11106.63	4919.571	19	0.0000	*MCM/GPCM*
20	400	80%	15848.42	10470.33	5378.092	19	0.0000	*MCM/GPCM*
20	400	90%	15241.45	9436.919	5804.526	19	0.0000	*MCM/GPCM*
20	1000	75%	40152.45	27716.3	12436.15	19	0.0000	*MCM/GPCM*
20	1000	80%	39998.02	26264.14	13733.88	19	0.0000	*MCM/GPCM*
20	1000	90%	38459.3	23457.42	15001.88	19	0.0000	*MCM/GPCM*
20	3000	75%	121392	83208.44	38183.53	19	0.0000	*MCM/GPCM*
20	3000	80%	118456.5	78960.23	39496.25	19	0.0000	*MCM/GPCM*
20	3000	90%	117017	70585.07	46431.94	19	0.0000	*MCM/GPCM*
40	400	75%	31563.14	21716.2	9846.939	39	0.0000	*MCM/GPCM*
40	400	80%	30434.34	20527.49	9906.847	39	0.0000	*MCM/GPCM*
40	400	90%	30492.42	18939.26	11553.16	39	0.0000	*MCM/GPCM*
40	1000	75%	79051.59	53770.54	25281.05	39	0.0000	*MCM/GPCM*
40	1000	80%	77270.35	51381.91	25888.45	39	0.0000	*MCM/GPCM*
40	1000	90%	76110.28	47044.5	29065.78	39	0.0000	*MCM/GPCM*
40	3000	75%	235818.2	161431.8	74386.37	39	0.0000	*MCM/GPCM*
40	3000	80%	234281.7	154481.4	79800.25	39	0.0000	*MCM/GPCM*
40	3000	90%	229789.3	141340.5	88448.79	39	0.0000	*MCM/GPCM*
60	400	75%	328589	202488.4	126100.6	59	0.0000	*MCM/GPCM*
60	400	80%	44666.06	30451.55	14214.51	59	0.0000	*MCM/GPCM*
60	400	90%	42868.17	27270.22	15597.95	59	0.0000	*MCM/GPCM*
60	1000	75%	77192.34	77192.34	33811.89	59	0.0000	*MCM/GPCM*
60	1000	80%	111052.9	74917.41	36135.53	59	0.0000	*MCM/GPCM*
60	1000	90%	107759.8	66914.78	40845.02	59	0.0000	*MCM/GPCM*
60	3000	75%	334728.6	232122.7	102605.9	59	0.0000	*MCM/GPCM*
60	3000	80%	336073.6	223638.5	112435.1	59	0.0000	*MCM/GPCM*
60	3000	90%	328589	202488.4	126100.6	59	0.0000	*MCM/GPCM*

In addition, the MCM/GPCM model provides higher value of test information in comparison to that of 3PLM/GRM model. The optimal value ratio of the test information from both scoring models for the 20-item, 40-item and 60-item respectively is 0.364583; 0.358974; 0.348485. In general, it can be stated that the function value of the test information which scoring made use of MCM/GPCM combination is three times higher than that of 3PLM/GRM. The finding supports the research that had been conducted and found by Donoghue (1994, p.300), Susongko (2009, p. 124) and Wasis (2009, p.105).

Finally, the answer for the main problem formulation is the summary of the first to the third problem formulation altogether with the results of the test on the model compatibiity to the data. The analysis of model compatibility test by means of -2LL statistic provides an MCM/GPCM model with better fit than the 3PLM/GRM model. The results support the findings of Chon, Lee, & Ansley (2007, pp.1-21). Therefore, in general, the MCM/GPCM combination is more superior in terms of interface in analyzing the mixed-item test format, especially in the Mathematics, than the 3PLM/GRM combination.

Conclusion and Suggestion

Conclusions

Based on the explanation on the results of the study, the researchers would like to draw the following five conclusions.

First, the combination of scoring model provides significant effect or influence in the level $\alpha = 0.05$ toward the test participants' θ ability estimates accuracy. The combination of MCM/GPCM model is more accurate than that of 3PLM/GRM model in estimating the θ ability. The bigger the sample size, the bigger the M-C item proportion and the longer the test length, the more accurate the θ ability estimates will be.

Second, the combination of MCM/GPCM scoring model has more accurate estimates on the item parameter than that of 3PLM/GRM model and the bigger the sample the more accurate the estimate results; however, the finding does not apply to the M-C test item proportion and the test length. Both by means of RMSE criteria and RMSD criteria, the estimates generated by both model combinations will be more accurate if the M-C test item proportion and the test length are smaller and shorter. In addition, the factors which have big influence are the model combination and the sample size while the M-C test item proportion has moderate influence. On the other hand, the test length does not have significant F value in the level $\alpha = 0.05$.

Third, in general, the researchers would like to state that the combination of MCM/GPCM model has provided the test information value three times higher than that of 3PLM/GRM model. In addition, for all of the test length the position of maximum test information value leads to the ability (θ) marginal estimates distribution. However, the researchers are unable to draw a 'solid' conclusion regarding the sample size and the M-C test item proportion in each test length that provided the optimum test information value.

Fourth, the θ ability standard estimates derivation error as well as the test parameter decrease under the estimation by means of MCM/GPCM combination in comparison to that of 3PLM/GRM. This finding implies that the MCM/GPCM scoring model is more accurate in estimating the θ ability and the test item parameter than the 3PLM/GRM is.

Fifth, the differences in the fit statistics between the two scoring models strengthen the superiority of MCM/GPCM combination upon the 3PLM/GRM combination at the level $\alpha = 0.05$.

Suggestions

The test developers, especially the ones who are responsible for the National Examination and the State University Admission Test, should consider the use of mixed-item test format in order to attain as much information as possible regarding the test participants' ability. In relation to the

matter, there should be considerations as well toward the wide-scale scoring implementation for the essay test items.

Then, the future researchers who would like to follow up the study are recommended to: (a) develop the model composition, for example the 3PLM/GRM combination, the MCM/GPCM combination and alike; (b) the numbers of response category in the study are made similar and there are four categories, therefore it is still possible that these categories might be developed into five categories or might be made different among the combined models because the researchers have not found the effects of the increase or the decrease on the model or even the unsimilarity of the response categories between the combined models; and (c) the criteria on the robustness test on the model during the unidimensionality assumption is violated because the data initiation for the IRT model combinations is assumed to be dimensional.

References

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale* (Unpublished doctoral dissertation). University of Massachusetts Amherst, USA. UMI Microform 9960735.

Berger, M. P. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement*, 22(3), pp. 248-258.

Bock, R. D. (1972). Estimating item parameters and latent ability when responsesare scored in two or more nominal categories. *Psychometrika*, 37(1) 29-51.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common item sets* (Unpublished doctoral dissertation). University of Maryland, Maryland USA.

Chon, K. H., Lee, W. C, & Anlsey, T. N. (2007). Assessing IRT model-data fit for mixed format tests. *CASMA Research Report*, Number 26

De Ayala, R. J. (1989). A comparison of the nominal response model and the three parameter logistic model in computerized adaptive testing. *Educational and Psychological Measurement*, 23(3), 789-805.

De Mars, C. E. (2008, March). *Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods*. Paper presented at the annual meeting of the National Council onMeasurement in Education, NewYork.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 31(4) pp. 295-311.

Ercikan, K. et al. (1998). Calibration and scoring oftests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, 35(2), pp. 137-154.

Garner, M., & Engelhard, Jr., G. (1999). Gender differences in performance on multiple-choice and constructed-response Mathematics items. *Applied Measurement in Education*, 12, pp. 29-51.

Gierl, M. J., Wang, C., & Zhou, J. (2008). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT. *Journal of Technology, Learning, and Assessment*, 6(6).

Glasersfeld, E.von. (1982). An interpretation of Piaget's constructivism. *Revue Internationale de Philosophie*, 36, pp. 612–635.

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Unpublished doctoral dissertation). University of Iowa, USA.

He, Y. (2011). *Evaluating equating properties for mixed-format tests* (Unpublished doctoral dissertation). University of Iowa, USA.

Hoskens, M. & De Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement*, 25, pp. 19-37.

Jurich, D., & Goodman, J. (2009, October). *A comparison of IRT parameter recoveryin mixed format examinations using PARSCALE and ICL.* Poster session presentedat the Annual meeting of Northeastern Educational Research Association, James Madison University.

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed response test scores: An economist's view. *Applied Measurement in Education*, 10, pp. 359-375.

Kentucky Department of Education. (2008). *Educational Planning and Assessment System (EPAS) College Readiness Standards and Program of Studies Standards Alignment Introduction* [Digital edition version]. Retrieved from http://www.education.ky.gov/

Kinsey, T. L. (2003). *A comparison of IRT and Rasch procedures in a mixed-item format test* (Unpublished doctoral dissertation). University of North Texas, USA. UMI Microform 3215773.

Lau, C. A. & Wang, T. (1998, April). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Levine, M. V., & Drasgow, F. (1983). The relationship between incorrect optionchoice and estimated ability. *Educational and Psychological Measurement*, 43, pp. 675-685.

Li, Y. H., Lissitz, R. W., & Yang, Y. N. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal Canada.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed-response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31, pp. 234-250.

Meng, H. (2007). *A comparison study of IRT calibration methods for mixed-formattests in vertical scaling* (Unpublished doctoral dissertation). University of Iowa, USA.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). New York: Pearson Education, Inc.

Sadler, P. M. (1998). Psychomatric models of examinee conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching,* 35(3), pp. 265-296.

Si, C. B. (2002). *Ability estimation under different item parameterization and scoring models* (Unpublished doctoral dissertation). University of North Texas, USA.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London: Academic Press.

Susongko, P. (2009). *Perbandingan keefektifan bentuk tes uraian dan testlet dengan penerapan 'graded response model' (GRM)* [The comparison between the effectiveness of explanatory test and test let with the implementation of graded response model] (Unpublished doctoral dissertation). Yogyakarta State University, Yogyakarta.

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with theone-parameter and two-parameter partial credit. *Journal of Educational Measurement*, 37, pp. 221-244.

Tall, D. O. et al. (2012). *Cognitive develop-ment of proof. In ICMI 19: Proof and Proving in Mathematics Education. Springer*. [Digital edition version]. Retrieved from http://homepages.warwick.ac.uk/staff/David.Tall/pdfs

Tang, K. L., & Eignor, D. R. (1997). Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models. *TEOFL Technical Report 13*. Princeton, NJ: Educational Testing Service.

Thissen, D. M. (1976). Information in wrong responses to the raven progressivematrices. *Journal of Educational Measurement*, 13(3), pp. 201-214.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501-519.

Thissen, D. M., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educationa Measurement*, 26(2), pp. 161-176.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice andconstructed-response tests. In R. E. Bennett, & W. C. Ward (Eds). *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrenc Erlbaum Associates.

Wainer, H. & Thissen, D. M. (1993). Combining multiple-choice and constructed response test scores: toward a marxist theory of test construction. *Applied Measurement in Education*, 6(2), pp. 103-118.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement*, 26(2), pp. 191-208.

Wasis. (2009). *Penskoran model partial credit pada item multiple true-false bidang fisika* [Partial credit scoring model on multiple true-false items in physics field] (Unpublished professor dissertation). Yogyakarta State University, Yogyakarta.