*International Journal of Instruction*
*e-ISSN: 1308-1470 ● www.e-iji.net*

*October 2020 ● Vol.13, No.4*
*p-ISSN: 1694-609X*
*pp. 545-558*

Received: 25/06/2019
Revision: 26/04/2020
Accepted: 16/05/2020
OnlineFirst:06/08/2020

# The Precision of Students' Ability Estimation on Combinations of Item Response Theory Models

**Ilham Falani**
Student, Departement of Research and Evaluation in Education, Universitas Negeri Jakarta, Indonesia, *ilhamfalani@gmail.com*

**Maruf Akbar**
Prof., Departement of Research and Evaluation in Education, Universitas Negeri Jakarta, Indonesia, *prof.dr.marufakbar@gmail.com*

**Dali S Naga**
Prof., Departement of Research and Evaluation in Education, Universitas Negeri Jakarta, Indonesia, *dalinaga@gmail.com*

This study compared the precision of ability estimation on different types of item response theory models for mixed-format data. Participants in this study were 1625 Junior High School Students in Depok, Indonesia. The mixed-format test was used to measure the students' ability in mathematics. The test used consists of multiple-choice and constructed response. Multiple-choice items are scored dichotomously, whereas constructed response items are scored polytomously. Furthermore, the mixed response data were analyzed using combinations of item response theory models. This study used a combination of Multiple-Choice Model for dichotomous data and Graded response model for polytomous data (MCM+GRM). Analysis of this model combination has never been done simultaneously. Test response data were analysed using PARSCALE. Furthermore, the estimation results were compared with the estimation results from a combination of 3 Parameters Logistic Model and Generalized Partial Credit Model (3PLM+GPCM). There are two criteria evaluation for the level of estimation precision: Root Mean Squared Error (RMSE) and correlation method. Based on the results obtained, the estimated RMSE value for the MCM+GRM is smaller than the estimated RMSE value with the 3PLM+GPCM. Also, the results of the estimated ability with MCM+GRM produce higher correlation values than 3PLM+GPCM. So, it can be concluded that the level precision of the MCM+GRM model is higher than 3PLM+GPCM. Therefore, MCM+GRM is more recommended for estimating students' mathematical ability in mixed-format tests.

Keywords: ability estimation, combination models, precision, combination, ability

**Citation:** Falani, I., Akbar, M., & Naga, D. S. (2020). The Precision of Students' Ability Estimation on Combinations of Item Response Theory Models. *International Journal of Instruction*, 13(4), 545-558. https://doi.org/10.29333/iji.2020.13434a

## INTRODUCTION

Mixed-format tests that include both multiple-choice (MC) and constructed-response (CR) items are widely used in large scale assessments (Bastari, 2000; Ercikan et al., 1998; Kim & Lee, 2004; Lissitz and et. al., 2012). It seems likely that the trend to combine items with different formats will be widely implemented. The MC and CR items are being frequently used in testing to complement each other to improved reliability, validity, and cost reduction. (Ercikan et al., 1998; Saen-amnuaiphon et al., 2012). Mixed-format tests can be challenging to analyze. Generally, MC and CR have different scoring schemas. MC items are dichotomously scored and CR items are polytomously scored (Alagoz, 2000; Kim & Lee, 2004). A technique is needed to analyze multiple scoring schemas in one set of test items. Classic test theory is difficult to apply to mixed-format tests because there is no model specifically designed to handle a combination of multiple scoring schemas in one set of test items (David et al., 1995; Kinsey et al., 2003). Model from item response theory (IRT) seems especially useful for analyzing dichotomous and polytomous response data simultaneously, as long as unidimensionality assumption holds. Simultaneous calibration may consist of either a mixture of different models or a single the simultaneous calibration is that it requires only a single run of an IRT estimate on program (Ercikan et al., 1998; Lee & Ansley, 2007)

The mixed-format tests consist of MC test items and CR test items that have been studied by several researchers. Previous studies revealed that current software packages can be used to analyze IRT models. Donoghue (1993) calibrated simultaneously the dichotomously scored items with 2 PLM and 3 PLM, and polytomously scored items with the GPCM using PARSCALE (Muraki & Bock, 1997). Bastari (2000) used MULTILOG (Thissen, 1991) and PARSCALE (Muraki & Bock, 1997) for analyzing the combination of 3PLM+GRM, and 3PLM+GPCM in order to estimate relationship in mixed-format test items in the common scale. The result shows that the combination of 3PLM+GPCM model provides higher precision ability estimation than 3PLM+GRM. Then, Abadyo (2015) continued a previous study to investigate the precision of ability estimation by using a combination of MCM for dichotomous and GPCM for polytomous, the result compared with 3 PLM+GPCM. The result shows that the combination of MCM+GPCM model provides higher precision estimation than 3PLM+GRM model within a mixed-format of mathematics tests. Based on the information presented above, further research is needed to compare the combined use of MCM+GRM and 3PLM+GPCM in increasing the precision of ability estimation.

### Purpose of The Study

The main objective of this study was to compare the precision of students' ability estimation on combinations MCM+GRM and 3PLM+GPCM.

## LITERATURE REVIEW

### Model Definition

For multiple-choice items or dichotomous scores, the model to be applied was the popular Birnbaum's three parameters logistic model or 3PLM model. This model was defined as follows:

$$P_j(\theta) = c_j + (1 - c_j)\left[1 + \exp\left(-Da_j(\theta - b_j)\right)\right]^{-1}$$

Thissen & Steinberg (1984), the Multiple Choice Model (MCM) is an extention of the Nominal Response Model (Bock, 1972). As an alternative approach to the limitation of the NR model, to handle the possibility that examinees with low proficiencies could choose any of the responses by guessing (Kim et al., 2002). The probability $P_{jk}$ under the MCM is expressed as

$$P_{jk}(\theta) = \frac{\exp[a_{jk}\theta + c_{jk}] + d_{jk}\exp[a_{j0}\theta + c_{j0}]}{\left[\sum_{h=0}^{K_j}\exp[a_{jh}\theta + c_{jh}]\right]},$$

$$\sum_{k=0}^{K_j} a_{jk} = 0,\ \sum_{k=0}^{K_j} c_{jk} = 0,\ \sum_{k=0}^{K_j} d_{jk} = 1,$$

For constructed-response items or polytomous scores, there are two kinds of models to be applied. One was Muraki's Generalized Partial Credit Model or GPCM (Muraki, 1992) while the other was Samejima's Graded Response Model or GRM (Samejima, 1969). GPCM is a common form of Partial Credit Model (PCM) (Muraki & Bock, 1997). GPCM is appropriate for analysis the successively ordered responses on a rating scale. GPCM is expressed in a mathematical form with the probability, $P_{ijk}$. The GRM is an extension of Thrustone's (1928) method of forming successive intervals to the analysis of graded responses on educational tests. Based on the assumption that an examinees' probability of scoring in score category k is described by the difference in probabilities of the person having scored greater or equal to k and having scored greater or equal to k+1. The GPCM and GRM models were formulated, respectively, below:

$$P_{jk}(\theta) = \frac{\exp\left\{\sum_{v=0}^{k} Da_j(\theta - b_{jv})\right\}}{\sum_{c=0}^{m_j}\exp\left\{\sum_{v=0}^{c} Da_j(\theta - b_{jv})\right\}}$$

$b_{j0} = 0.0.$

and

$$P_{jk}(\theta) = \frac{1}{\left[1 + \exp\left(-Da_j(\theta - b_{jk})\right)\right]} - \frac{1}{\left[1 + \exp\left(-Da_j(\theta - b_{j,k+1})\right)\right]},$$

$$P_{jk}(\theta) = P^{+}_{jk}(\theta) - P^{+}_{j,k+1}(\theta),$$

$$P^{+}_{j0}(\theta) = 1.0, P^{+}_{j,K+1}(\theta) = 0.0.$$

with,

$P_j(\theta)$ = probability of correctly responding item j;
$P_{jk}(\theta)$ = probability of responding category *k* on item *j*;
$\theta$ = proficiency level;
$a_j$ = discrimination (slope) parameter for item *j*;

$b_{jk}$ = threshold or step parameter for item *j* on category *k*.
*cj* = guessing parameter for item j.
$d_{jk}$ = proportion of those who don't know for item j on category k.
*D* = scaling factor (typically 1.7);
*k* = number of categories;

**Ability Estimation**

The ability estimation with the item response theory is done by using the items that have been calibrated. The items in this test are considered to have discrimination (slope), threshold, and guessing. Various techniques and approaches can be applied to estimate abilities in IRT. The ability estimation can be carried out separately with item parameters estimated in advance, or simultaneously with the item parameters (Naga, 2012). If the ability parameter is not estimated along with the item parameter, the item parameter is first estimated from the item response by eliminating the effect of the ability parameter, the ability parameter can be eliminated through conditioning or integrated through marginalization (Si & Schumacker, 2004). Techniques that can be used in parameter estimation, namely: minimum chi-quadrant (Zwinderman & Arnold, 1983), Bayesian capital estimation procedure (Frank, 1992; Mislevy, 1986), logistic regression (Reynolds et al., 1994), and maximum likelihood procedure (Baker, 1992). In this study, ability estimation is done by using the maximum likelihood procedure.

**Maximum Likelihood Estimation**

The ability estimation can be done using the Likelihood function. The probability of examinee with $\theta$ ability, giving a response to an item, is expressed by (Hambelton & Swaminathan, 1985).

$$U_i \begin{cases} 1 & \text{for correct answers} \\ 0 & \text{for wrong answers} \end{cases}$$

For correct answers, the probability can be written $P(U_i = 1|\theta)$, this probability is a function of response items, which are usually written $P_i(\theta)$ or $P_i$. The probability of a response can be expressed by,

$$P(U_i|\theta) = P(U_i = 1|\theta)^{U_i}\, P(U_i = 0|\theta)^{1-U_i}$$

$$P(U_i|\theta) = P_i(1-P_i)^{1-U_i}$$

$$P(U_i|\theta) = P_i(Q_i)^{1-U_i}$$

If the examinee with $\theta$ ability respond to *n* items, the probability of responses $U_1, U_2, ..., U_n$ expressed by $P(U_1, U_2, ..., U_n|\theta)$,

$$P(U_1, U_2, ..., U_n \mid \theta) = P(U_1 = 1 \mid \theta) P(U_2 = 1 \mid \theta), ..., P(U_n = 1 \mid \theta)$$
$$= \prod_{i=1}^{n} P(U_i \mid \theta)$$
$$= \prod_{i=1}^{n} P^{U_i} (Q_i)^{1-U_i}$$

The equation above is a joint probability for $n$ items. However, when the response is observed, for example, random variables $U_1, U_2, ..., U_n$ , has a specific value $u_1, u_2, ..., u_n$ , the value is 0 or 1, then the equation above is no longer probability. The equation becomes mathematics function called Likelihood Function, and expressed by (Hambleton & Swaminathan, 1985),

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{i=1}^{n} P^{U_i} (Q_i)^{1-U_i}$$

With $\theta$ follow

$$\frac{d}{d\theta} \ln L(u \mid \theta) = 0.$$

The equation is above known as the probability equation. This equation is non-linear and cannot be resolved explicitly.

The common method used to solve these equations is a numerical approach using the New-Rapshon procedure. In this study, ability estimation was carried out using PARSCALE software. PARSCALE is a calibration program that uses the Maximum Likelihood Estimation (MLE) procedure to estimate items and ability parameter in 1, 2, and 3 parameters logistic models, as well as some polytomous models such as (PCM, GPCM, GRM, GRSM) (Ayala, 2013).

**Evaluation Criteria**

Two criteria are used to determine how close the ability estimation results to the true ability, first, Root Mean Square Error (RMSE) of the estimated ability (Si & Schumacker, 2004). RMSE shows bias and ability estimation variants so that it serves as a precision indicator. RMSE calculation is carried out on the estimation results of the two combination models used, then the RMSE value is compared to see which combination of models results in an estimation of higher precision levels. The second criterion is the most common one, the correlation method (Bastari, 2000). Correlation between true abilities and their estimates were calculated. The results then were averaged across replications. RMSE and the correlation value is calculated by the formula, respectively,

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n} \left(\bar{\theta}_j - \theta_j\right)^2}{n}}.$$

$$\text{Correl}\left(\theta,\bar{\theta}\right)=\frac{\sum_{j=1}^{n}\left(\theta j-\bar{\theta}\right)\left(\bar{\theta}\, j-\bar{\bar{\theta}}\right)}{\sqrt{\sum_{j=1}^{n}\left(\theta j-\bar{\theta}\right)^{2}\sum_{j=1}^{n}\left(\bar{\theta}\, j-\bar{\bar{\theta}}\right)^{2}}}.$$

Where $n$ is the number of examinee, $\theta_j$ is the true ability for the examinee j. $\bar{\theta}$ is the average of estimated abilities. Whereas $\bar{\bar{\theta}}$ is the average of the estimated abilities of participants from 10 replications.

## METHOD

### Participants and Procedures

Participants in this study were 1625 eighth grade students enrolled in eight public schools in Depok, Indonesia. These schools were selected based on the highest level of school accreditation. The Participants consisted of 995 girls and 630 boys. It is intended that selected participants can answer the test properly. The estimation of mathematical abilities carried out on the participants of this study.

The research data collection is divided into two stages. The first stage is a trial test, this trial is conducted to calculate the value of the validity and reliability of the test. The second stage is the collection of research data using tests that have been improved based on the first stage. The duration of the data collection stage lasts approximately two months. Collecting data begins with a research permit application to the school to be addressed, after obtaining permission from the school. With the help of the mathematics teacher for each class, the researcher disseminated tests in the field of mathematical studies that had been prepared for students to do. Then the student test results data are collected and conducted scoring.

Scoring for multiple-choice is done in two ways. The first, multiple-choices scored with the Polytomous model, provided that the further the relationship between the choice of answers and the answer key, the smaller the score given. Scores given are 1, 2, 3 and 4. Example:

The function is defined as f (x) = –12 + 10x. The value of f (–2) is ...

(Answer key: -32)

| | | | |
|---|---|---|---|
| A. | –32 (score 4) | C. 8 | (score 1) |
| B. | –8 (score 2) | D. 32 | (score 3) |

The second method, converted previous polytomous scoring model to dichotomous model, for each question given a score of "0" if the answer is wrong or "1" if the answer is correct. This means score "1" for the answer that matches the key, score "0" for another answer.

Scoring for the construct response is done by the Polytomous model, with the following conditions: score "0" if there is no correct step, score "1" if the first step is correct, score

"2" if until the second step is correct, score "3" if until the third step is correct, and the score is "4" if all steps are correct.

The response data that has been scored is then inputted into the Microsoft Excel format for further analysis for the estimation of the mathematical ability of the examinees'. This data collection was carried out for three months, considering that data collection was carried out in eight different places and different licensing procedures.

**Instrument Test**

The instrument test used in this study is a mixed format test consisting of multiple-choice items and essays. The test consists of 30 multiple-choice items and 5 construct responses used in this study. The multiple-choice items arranged have four answer choices for each question. As for the CR item, the answer format is set to consist of four answer steps. This is to facilitate the scoring schema. The preparation of questions is based on the Minister of Education and Culture Regulation of the Republic of Indonesia No. 20 of 2016 concerning Competency Standards for Primary and Secondary Education Graduates. These competency standards are translated into basic competencies in each subject area. This study focused on the field of mathematics for the first semester. Furthermore, these basic competencies are described in each indicator, where each indicator represented by at least one question. To ensure the validity of the contents of the test that has been prepared, the contents of the test are validated by experts who have a background in mathematics, namely three lecturers and five junior high school mathematics teachers. Calculation of content validity by experts was carried out using the Aiken Method. After going through the process of content validity by experts, then the test instrument was tested right to calculate the level of reliability and validity of the test. The trial of this test instrument was carried out once before this test was used for research data collection. This trial aims to calculate the statistical value of validity and reliability tests. Test items were improved based on the results of the first trial. Furthermore, the improved test is used for research data collection.

**Data Analysis**

Collected responses of the examinees then scanned, recapitulated, and analyzed. The analysis carried out parameter estimation based on the results of the response data. Table 1 presents the scoring combination and model analysis response data for each of the data from the results of the examinees' responses.

Table 1
Scoring Combination and Model Analysis Response Data

| Format Item | Number of items | Data I | | Data II | |
|---|---|---|---|---|---|
| | | Scoring Format | Analysis Model | Scoring Format | Analysis Model |
| MC | 30 | Dichotomous | 3PLM | Polytomous | MCM |
| CR | 5 | Polytomous | GRM | Polytomous | GPCM |
| Analysis | | Simultaneously | | Simultaneously | |

The data in this study consists of two types of data that come from two different scoring formats, as in the table above. The next step is to perform simultaneous calibration of two items formats simultaneously. This study only focused on parameter estimation of examinees' ability. The two combinations of models used to analyze Data I and II respectively are a combination of 3 PLM and GPCM (3PLM+GPCM) and a combination of MCM and GRM (MCM+GRM).

For each data, an estimate of the ability of the examinees carried out using Marginal Maximum Likelihood (MML) estimation. Furthermore, the marginal mean estimation was compared. This is done to see the estimated Precision of the two-model combinations. The analysis was carried out using Bilog MG software (Zimowski et al., 1996). Therefore, it is necessary to synthesize a combination of IRT models for the analysis of mixed-format tests on BILOQ MG. Data from the estimation of this ability will be considered a "true ability" for each examinee (Susan, 1996). To investigate the estimation precision-made, then the response data replication was 10 data with the true ability and grain characteristics of each initial data. Replication of the response is done using software Wingen. From the 10-replication data available, it will then be correlated with the true ability. A high correlation value will show which model has higher precision in estimating the ability of examinees.

**FINDINGS**

Based on the results of participant tests in this study, mixed data were obtained consisting of dichotomous and polytomous. Data analysis was performed using a combination of IRT models. Furthermore, students' mathematical ability (theta) are estimated based on these data. Estimates are carried out simultaneously.

Before analyzing the data with PARSCALE, unidimensionality assumption was tested by performing factor analysis of data (Alagoz, 2000). The LISREL 8.5 software (Joreskog & Sorbom, 2003) run indicated that most items loaded on a single factor possibly called the general mathematics ability. The results from factor analyses show that the principal axis factoring extracted 2 factors with the eigenvalues greater than 1. The first and second factors explained respectively, 43,27% and 5,7% of the total variance. Meanwhile, the other factors explained ranging from 0.25 to 0.65% of the total variance. This result indicated that data seems to be reasonably unidimensional.

Furthermore, data generated from different scoring schemes (Table 1) are analyzed using a combination of IRT models, namely: 3PLM+GPCM and MCM+GRM. Analysis with the IRT model was carried out with PARSCALE. The normality test is done using MINITAB 17. The results of the test showed that the θ distribution was not normal. The process of editing data is done by reducing some of the extreme scores, then the distribution normality test is repeated until the distribution of the ability of the examinee to the estimation results for both scoring schemes is normal.

Table 2
Descriptive Statistics of True Theta

| Var. | N | Mean | SE Mean | StDev | Variance | Min | Max | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| Data I | 1531 | -0,000 | 0,0254 | 0,9931 | 0,9863 | -2,7003 | 3,3006 | 0,02 | 0,13 |
| Data II | 1531 | -0,000 | 0,0245 | 0,9586 | 0,9190 | -3,9795 | 3,7719 | -0,22 | 0,61 |

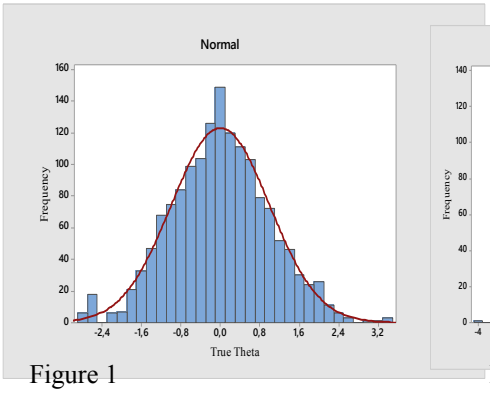


Figure 1
Histogram of True Theta 3 PLM+GPCM

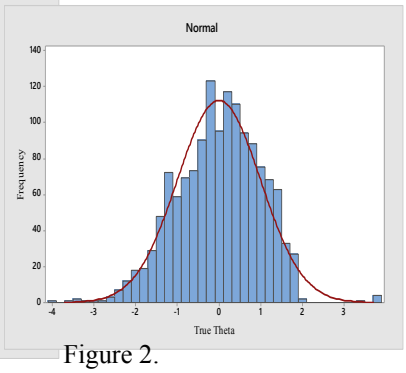


Figure 2.
Histogram of True Theta MCM+GRM

The normality test is done by the Anderson-Darling method. Normality test in Data I and II are done with the provisions if the P-Value is more than 0.05, the data is normally distributed. Otherwise, if the P-Value is less than 0.05 then the data is not normally distributed (Darling, 2015). P-Value for Data I and II respectively 0.150 and 0.06. Both of the P-value is more than 0.05, so it can be concluded that the Data I and II are normally distributed.

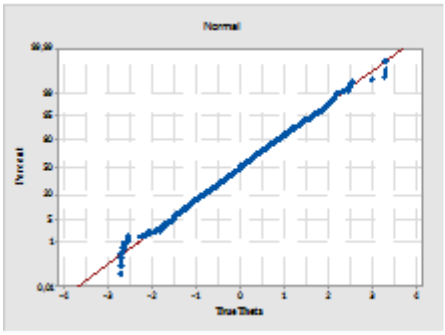


Figure 3
Probability Plot of True Theta Data I

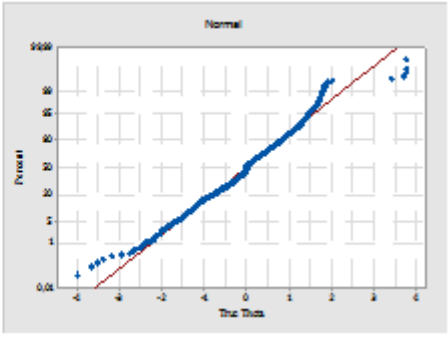


Figure 4
Probability Plot of True Theta Data II

The estimation results of examinees' ability to Data I and II produced were used as a "true theta", to replicate the data response 10 times. This was done to compare the evaluation criteria used in this study.

Here is a comparison of the results of the calculation of the RMSE (evaluation criteria 1) the estimation of examinee's ability to use both the combination of IRT models used,

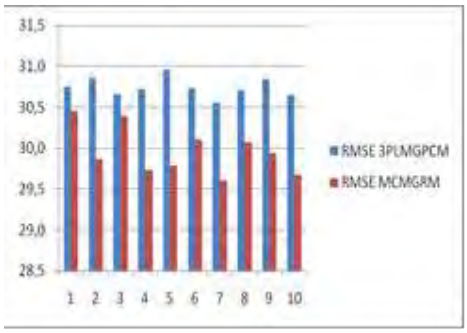


Figure 5
Comparison of RMSE

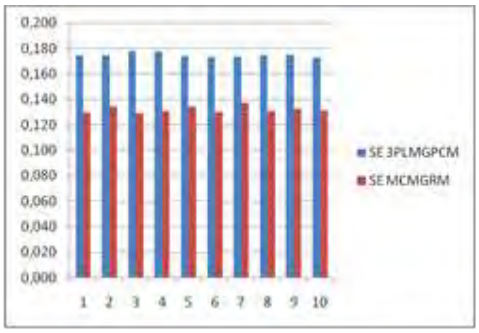


Figure 6
Comparison of Standard Error Estimation

Based on Figure 5, it can be seen that from 10 data replications, RMSE values for the combination of MCM+GRM models are smaller than 3 PLM + GPCM combination. Also, Figure 6 showed the standard error of estimation is obtained from PARSCALE output, it is known that the standard error of estimation value for MCM+GRM is smaller than 3 PLM + GPCM. This occurred in 10 replication data.

The following is the comparison of the correlation results (Evaluation criteria 2) in estimating examinees' ability by using both combinations of IRT models used,

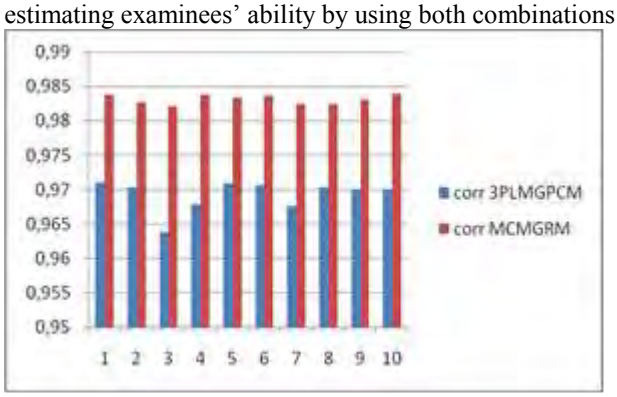


Figure 7
Comparison of Correlation Estimation

Figure 7 shows the comparison of the correlation between the 10 data replications and true ability, the result that MCM+GRM has a higher correlation than the 3PLM+GPCM. This shows that the combination of MCM+GRM is more superior in the evaluation criteria 2.

Table 3
Descriptive Statistics for the Ability Estimation Errors

| Mixture Model | SE | RMSE | Correlation |
|---|---|---|---|
| 3PLM+GPCM | 0,175 | 30,700 | 0,961 |
| MCM+GRM | 0,130 | 29,500 | 0,983 |

Can be seen in the table above is based on two evaluation criteria used the RMSE and correlation. The combination of the MCM + GRM model produces better values than the comparison combination.

## DISCUSSION

Based on the results obtained, it can be concluded that the format of scoring items affects the level of estimation precision. The ability estimates resulting from polytomous scoring had slightly higher measurement precision than those resulting from dichotomous scoring (Jiao & Liu, 2014; Fung, 2002). Polytomous IRT modeling can result in more precise estimates of examinee ability (Bolt et al., 2001; Baker, 1992). It was previously mentioned that the MCM+GRM scoring format was carried out with the Polytomous model for both MC and CR, while the 3 PLM+GPCM with the combined Dichotomous and Polytomous models.

RMSE was a measure of the average deviation of the estimators from the known ability estimates. The results showed that the combination of the MCM + GRAM model produced an estimated RMSE value that was smaller than the combination of the comparison models. The results indicated that MCM+GRM produced more precise estimates than 3 PLM+GPCM under all replications. This is also supported by the results of the second evaluation criteria which show that MCM + GRM produces higher correlation values. This shows that the combination of MCM + GRM model is superior.

The results of this study can be a reference for teachers in selecting the right combination of models to measure the ability of students. Precision measurement results in a good assessment. Furthermore, the results of this assessment will be an evaluation material for a teacher to improve the quality of learning undertaken. If the evaluation material used does not reflect the actual condition, then the corrective steps to be taken may not be appropriated. Therefore, the teacher needs to be able to measure students' ability precisely.

This study has limitations, the sample size and fixed test length. For further investigation about the estimation precision in MCM + GRM, a simulation study can be conducted. Simulation studies can produce more varied conditions so that we can find out the effect of sample size, test length, and proportion of test items on ability estimates on the combination of MCM + GRM models.

## CONCLUSIONS

This research compared the precision of IRT model combinations for mixed-format data. The estimation results carried out simultaneously using PARSCALE show that the combination of the MCM + GRM model produces a smaller RMSE value compared to 3 PLM + GPCM. The average correlation value between the estimation and replication results for the MCM + GRM combination is higher compared to 3PLM + GPCM. Based on the two evaluation criteria, it was concluded that the combination of the MCM+GRM model resulted in a more precise estimation of the ability of students compared to 3PLMG+PCM. Therefore, the MCM + GRM combination is recommended for estimating students' ability on mixed-format tests. Precise estimation results greatly

affect the assessment to be carried out by the teacher, a good assessment will be very useful for teachers in conducting evaluations and improving the quality of learning. For further research, it is recommended to investigate the estimation of students' abilitiy on the combination of the IRT model on mixed-format tests based on sample size, test length, and proportion of test items.

**ACKNOWLEDGMENTS**

We sincerely thank to LPDP (Lembaga Pengelola Dana Pendidikan), Ministry of Finance, Republic of Indonesia for funding this research.

**REFERENCES**

Abadyo, & Bastari. (2015). Estimation of ability and item parameters in mathematics testing by using the combination of 3PLM/GRM and MCM/GPCM scoring model. *Research and Evaluation in Education Journal*, *1*(1), 55–72.

Alagoz, C. (2000). *Scoring tests with dichotomous and polytomous* (Unpublished master thesis). University of Georgia, Georgia.

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data, *Journal of Educational and Behavioral Statistics*. *26*(4), 381–409.

Darling, D. A. (2015). A test of goodness of fit. *Journal of the American Statistical Association*. *49*, 765–769.

Bock, R. D (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. doi:10.1007/BF02291411.

Chon, H. K., Lee, W. C., & Ansley, T. N. (2007). *Center for advanced studies in measurement and assessment Casma research report assessing IRT model-data fit for mixed-format tests* (Report No. 26). Iowa City: University of Iowa.

David, T., Pommerich, M., Kathleen, B., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49.

Ayala, R. J. (2013). *The theory and practice of item response theory*. New York: Guilford Publications.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*(2), 137–154. doi: 10.1111/j.1745-3984.1998.tb00531.x.

Frank B, B. (1992). *Item response theory: The basics of item response theory*. USA:

ERIC

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer. doi:10.1007/978-94-017-1988-9.

Jiao, H., Liu, J., Haynie, K., Woo, A., & Gorham, J. (2014). Comparison between dichotomous and polytomous scoring of innovative items in a large-scale computerized adaptive test. *Educational and Psychological Measurement*, *72*(3), 493-509. doi: 10 .1177/0013164411422903.

Donoghue, J. R. (1993). *An empirical examination of the IRT information in polytomously scored reading items* (Report No. RR-93-12). New Jersy: Educational Testing Service.

Kim, J., Madison, W., Hanson, B. A., & Mcgraw-hill, C. T. B. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement. 26*(3). 255–270.

Kim, S., & Lee, W. (2004). I*RT Scale linking methods for mixed-format tests* (Report No. 5). Iowa: ACT Research.

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

Lissitz, Robert, Xiadong Hou, & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*(3), 1-50.

Lord, F.M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi: 10.1177/014662169201600206.

Muraki, E., & D. Bock, R. (1997). *Parscale: IRT item analysis and test scoring for rating scale data*. Chicago: Scientific Software International.

Naga, D. S. (2012). *Teori Sekor pada Pengukuran Mental*. Jakarta: PT Nagarani Citrayasa.

Reynolds, T., Perkins, K., & Brutten, S. (1994). *A comparative item analysis study of a language testing instrument. Language testing*. New York: Sage Publications. doi: 10.1177/026553229401100102.

Saen-amnuaiphon, R., Tuksino, P., & Nichanong, C. (2012). The effect of proportion of mixed-format scoring: Mixed-format achievement tests. *Procedia - Social and Behavioral Sciences*, *69*, 1522–1528. doi: 10.1016/ j.sbspro. 2012.12.094.

Si, C. F., & Schumacker, R. E. (2004). Ability estimation under different item parameterization and scoring models. *International Journal of Testing*, *4*(2), 137–181. doi: 10.1207/s15327574ijt04023.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory*. New York: Scientific Software International.

Thissen, D., & Steinberg, L. (1984). A response model for multiple-choice items. *Psychometrika*, *49*(4), 501–519. doi: 10.1007/BF02302588.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [computer program]*. Chicago: Scientific Software International.

Zwinderman, A. H., & Arnold, L. (1983). Robustness of marginal maximum likelihood estimation in the Rasch Model. *Applied Psychological Measurement, 14*(1), 73–81.