academicJournals

Vol. 11(16), pp. 1579-1589, 23 August, 2016
DOI: 10.5897/ERR2016.2807
Article Number: B2D124860158
ISSN 1990-3839

http://www.academicjournals.org/ERR

Educational Research and Reviews

Full Length Research Paper

Variance difference between maximum likelihood estimation method and expected A posteriori estimation method viewed from number of test items

Jumailiyah Mahmud*, Muzayanah Sutikno and Dali S. Naga

[1]Institute of Teaching and Educational Sciences of Mataram, Indonesia.
[2]State University of Jakarta, Indonesia.
[3]University of Tarumanegara, Indonesia.

Received 8 April, 2016; Accepted 12 August, 2016

The aim of this study is to determine variance difference between maximum likelihood and expected A posteriori estimation methods viewed from number of test items of aptitude test. The variance presents an accuracy generated by both maximum likelihood and Bayes estimation methods. The test consists of three subtests, each with 40 multiple-choice items of 5 alternatives. The total items are 102 and 3159 respondents which were drawn using random matrix sampling technique, thus 73 items were generated which were qualified based on classical theory and IRT. The study examines 5 hypotheses. The results are variance of the estimation method using MLE is higher than the estimation method using EAP on the test consisting of 25 items with F= 1.602, variance of the estimation method using MLE is higher than the estimation method using EAP on the test consisting of 50 items with F= 1.332, variance of estimation with the test of 50 items is higher than the test of 25 items, and variance of estimation with the test of 50 items is higher than the test of 25 items on EAP method with F=1.329. All observed F values ≥ 1.00. 5 RMSE in items 10, 15, 20, and 25 are different in both MLE and EAP, with t = 3.060, $\alpha = 0{,}011$**, thereby meaning that statistical null hypothesis are rejected. The study concludes that variance of MLE method is higher than EAP, and the test with 50 items has higher variance than that with 25 items, the accuracy of EAP estimate higher than that of MLE in item 10, 15, 20, and 25.**

Key words: Variance, RMS of estimation, maximum likelihood, expected A posteriori.

INTRODUCTION

There are two types of psychological and educational measurement theories; classical and modern. Such a modern measurement theory is also known as item response theory which is developed in response to the weakness of the classical measurement theory, mainly in its dependence among groups of test-takers and items. Dependence means that result of measurement depends on groups of those who do the test. If such a test is

*Corresponding author. E-mail: jumailiyah@gmail.com.

provided to groups of its takers who have high proficiency, level of difficulty of items assessed in the test is getting lower. On contrary, if it is given to those with low capabilities, level of difficulty of the test items becomes higher (Hambleton, 1991). It was found that the classical test theory (CTT) had some limitations, however, Item Response Theory (IRT) showed a variety of benefits such as:

1. Estimating item difficulty
2. More stable in terms of difficulty indeces
3. More stable of internal consistency, and
4. Markedly reducing in error's measurement (Magno, 2009).

IRT, as a model- and item-basedapproach, is obviously considered succesful in its use in terms of research and practice applications. The main role of IRT model is to estimate someone's position on a latent dimension (Reise and Revicki, 2014). In addition, estimation quality depends on accuracy criteria consisting mean square error (MSE), bias and estimation variance. The results showed that parameter estimation is much more better by implementing priory, particularly for two and three parameter models (Baker, 2004).

The aim of educational measurement is to know level of test takers' ability that could be used in selections for decision-makings. Results of such selections are used to identify whether or not candidates can be accepted in a particular program. A decision to accept the candidates or not is often wrong which may bring negative implications to further individuals' developments. An inappropriate decision is often caused by the use of invalid and unreliable instruments or tests, results of measurements which are very different from the actual conditions so that it contains high uncertainty. In contrast, a measurement is believed containing high accuracy if its result has small RMS and variance.

Perspective proposed by DeMars reveals assumptions on item response theory relating to unidimensional, local independency and model specification accuracy (DeMars, 2010). Firstly, unidimensional is a test measuring only a character or a particular test takers' ability. Items in a test, for example, just measure participants' ability to count, not to assess their proficiency in a language either. Statistically, unidimensional can be calculated its Eigen score using factor analysis, indicated by a dominant one. Secondly, local independency is meant that the influence of test takers' ability and test items is supposed to be constant; test takers' responses on the items are not statistically connected. "This assumption can be accepted if test takers' answer in a certain item do not influence answers in other items. Test takers' answers in some test items are expected to be unrelated" (Hambleton et al., 1991). Implication of such assumption causes is that the test can be analyzed item per item. For the test takers, an analysis is also done on individual basis.

Correct answer probability, item parameter and test takers' characteristics are correlated in a logistic formula model. As a result, item curve characteristic is reflected by logistic model used as a basis of calculation (Naga, 1992). Model description of a parameter in a curve of item characteristic is level of difficulty of item itself (b_i). The higher level of difficulty a certain item has, the higher ability the test takers need to answer such item correctly. Thus, as shown in the location, the higher level of difficulty the test item has, the righter position it will be.

In a two-parameter logistic model, it shows that such a model calculates item level of difficulty (b_i) and item discrimination index (a_i). The picturing in item arch curve (ICC), item discrimination index is shown in the curve of item with slope or curve precipitousness. An item with high precipitousness shows high discrimination index or value of a_i is high. In contrast, slope item curve shows low discrimination index or value of a_i is low. Here is a formula for two-parameter logistic model:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \text{ i=1,2,3,} \quad (1)$$

Pi (θ) = probability of test takers having ability θ chosen randomly that could answer item i correctly.
θ = subject's level of ability
bi = parameter of item level of difficulty for i
ai = discrimination index for item i
n = number of items in a test
e = numeral valuing 2.718
D = scaling factor made in order that logistic function closes to function of normal ogive (Hambleton et al., 1991)

Estimation is means fathoming or apprising. Estimation contains the findings of appropriate values for parameters of equality using certain ways or methods (Makridakis et al.,1999). Some differences, however, may be found between them. Such differences are:

1. Regression model is often applied in variables that have linear correlation; while in parameter logistic model, there is a correlation between test item and nonlinear test takers' ability.
2. Independent variable in regression model is observable. On contrary, independent variable of test takers' ability $(\theta)$is unobservable in item response theory (Hambleton et al., 1991). Since actual scores of test item parameters and ability of test takers are unidentified, an analysis and estimation for parameters of test takers' ability and item scores is required to do. The identification of scores of parameters is known as parameter estimation.

A study by Borgatto and Pinheiro (2015) aims to determine the impact of ability estimation on IRT with respect to the difficulty level of the test and to evaluate

whether the error of estimation can be influenced by method of estimation used. There are 2 estimation methods used, which are weighted likelihood estimator (WLE) and MLE. EAP method is divided according to its prior θ uniform and normal distribution, and this uniform distribution is compared with WLE. The standard measurement to determine the accuracy of estimation method is Mean Squared Error (MSE). Simulation uses 4 tests based on the number of items which are 15, 30, 45 and 60 items. The result of study showed that uniformed EAP has better MSE than WLE with 45 items, and from this result it was suggested to use more items in estimating ability parameter.

Accuracy is meant as the most criteria used to evaluate the works of models of estimating methods. Accuracy shows level of correctness of estimating result. It can be measured using average dimension of square error which is referred to as *Mean Square Error* (Makridakis et al., 1999). A standard error of estimate is commonly composed of two components, including bias in the estimate and MSE (de Ayala, 2009).

Uncertainty of measurement contains measurement estimation error. In a regression method, correlations of two variables stated in simple regression lines in forms of justifiable and unjustifiable errors. An estimation of student's scores in a particular test is done using logistic analysis which is basically in line with what Hambleton (1991) stated that logistic model is nonlinear correlation and measuring invisible matter. Although logistic model is nonlinear correlation, the interpretation of fault or *error* is equal to deviation. Total deviation $(Y_i - \bar{Y})$ consists of unjustifiable deviation $(Y_i - \hat{Y})$ and justifiable deviation $(\hat{Y}_i - \bar{Y})$ (Makridakis et al., 1999).

Justifiable deviation shows differences between result of ability estimation and its mean score $(\hat{Y}_i - \bar{Y})$, and this was used as a basis for calculating estimation variance in this research analysis discussion. As previously discussed, the estimation accuracy is indicated by *Mean Square Error*, bias, and variance. Such variance is seen as a part of concept of measurement uncertainty. Variance (σ^2), in which the high one shows high uncertainty in its parameter measurement. Variance also shows score distribution. The bigger score range distribution it has, the higher variance it will be. This means that the accuracy of estimation measurement is low or vice versa. Score of parameter obtained through such estimation contain big variance. Thus it is not very accurate. In contrast, if score of parameter obtained through estimation contain small variance, it $\sum s^2$means that the score of parameter is sharp and accurate (Naga, 1992).

Estimation accuracy with *residual error* score is also referred to as *mean squared error* (MSE), while variance is indicated by a distinction between estimation score and average score $(\hat{Y}_i - \bar{Y})$ as a justifiable deviation. Hambleton et al. (1991) explains that score of *residual error* of *maximum likelihood estimation* method is higher than that of *expected a posteriori* method. Hambleton et al. (1991) clearly states that scores of *residual error* of MLE is higher than EAP. This indicates that the accuracy of EAP estimation is better than MLE (Hambleton et al., 1991). Swaminathan et al. (2003) argues that "*Bayes* procedure generally can result to smaller variance compared to that generated from *maximum likelihood*. Baker et al. (2004) points out an estimation bias resulted from both of the estimation methods; "*there was little difference between the MLE/EM estimates and those obtained via the Bayesian procedures*". This means that accuracy of both estimation methods is equal or at least it has a little difference so it is possibly ignored. There are many items in a set of test in *item response theory*, and test length estimation variance influence accuracy of ability estimation.

In a classical testing theory, longer tests are more trustful than those of the shorter ones, yet in an item response theory, shorter test may be more reliable than those of the longer ones. This can be seen in a *computer adaptive test* in which level of difficulty is adjusted with test takers' ability and will bear small measurement error (Embretson and Reise, 2000).

The accuracy of ability estimation is not visible when using a few numbers of items because it requires many of those in order to be able to judge it. A test with 30 items indicates fixed error much lower than that with 20 items (Embretson and Reise, 2000). Referring to experts' points of view (Embretson and Reise, 2000):

1. 30 items indicate fixed error smaller in number than that with 20 items. This is in line with classical theory in which it believes that the more items in a package of test, the more trustful it will be
2. In item-response theory, there is no guarantee for the greater number of items in a test package to have little number of errors when comparing to that with small number of items.

Referring to the theoretical discussion, some hypotheses can be addressed as follows:

1. Variance of ability estimation method (*θ)* of MLE is higher than that of ability estimation (*θ)* of EAP in a test with 25 items.
2. Variance of ability estimation method (*θ)* of MLE is higher than that of ability estimation (*θ)* of EAP in a test with 50 items.
3. Variance of ability estimation *(θ)* of a test consisting 50 items is higher than that of a test with 25 items in MLE method.
4. Variance of ability estimation *(θ)* of a test consisting 50 items is higher than that of a test with 25 items in EAP

method.
5. Accuracy of RMS in EAP method is higher than that of MLE method particularly in item 10, 15, 20 and 25.

METHODOLOGY

Procedure of ability estimation can be done using method of *Maximum Likelihood (ML), MAP and EAP* (Embretson and Reise, 2000). The word *"likelihood"* interpreted as possibility or probability' while *"maximum"* means big opportunity. "Maximum likelihood", therefore, may be interpreted as probability that has biggest opportunity. "*Maximum likelihood"* is a model of "*total likelihood*" (Du Toit, 2003). This biggest opportunity will depend on probability of correct and wrong answer made by test takers when doing a particular test as well as logistic model used in it. To indicate maximum score, calculation of iteration is done (Baker and Kim, 2004). Ability estimation of *maximum likelihood* method determines score of maximum ability belongs to each test taker, calculating process of formula 1 up to 5 (Du Toit, 2003). Score of $P_j(\theta)$ acquired from formula 1, $L_i(\theta)$ from formula 2 in term of multiplication, notated in *quadrature* $P_j(\theta) = P(X_k)$ as follows:

$$\log L_i(\theta) = \sum_{j=1}^{n}\{x_{ij}\log_e P_j(\theta) + (1 - x_{ij})\log_e[1 - P_j(\theta)]\} \quad (2)$$

$L_i(\theta)$ = Score of maximum ability for each test taker.

$P_j(\theta)$= Probability of ability in an item as shown in formula 1.

x_{ij} =Number of correct item.

1-x_{ij}= Number of wrong item.

Maximum score of ability of each test taker $L_i(\theta)$ derived in logarithm which is equal to null using the following formula:

$$\frac{\partial \log L_i(\theta)}{\partial \theta} = \sum_{j=1}^{n} \frac{x_{ij} - P_j(\theta)}{P_j(\theta)[1 - P_j(\theta)]} \frac{\partial \log(\theta)}{\partial \theta} = 0 \quad (3)$$

The estimation of ML,θ is calculated using Fisher scoring which is commonly known as "*Fisher information*". Formula used for two-parameter model is as follows:

$$I(\theta) = \sum_{j=1}^{n} a_i^2 P_j(\theta)\left[1 - P_j(\theta)\right] \quad (4)$$

I(θ) = information function of respondent's ability.
θ = respondent's level of ability.
a_i = discrimination index of item i.

After obtaining score of ability information function from a two-parameter logistic model, an iteration is done using the following formula:

$$\theta_{t+1} = \theta_t + I^{-1}(\theta)\left(\frac{\partial \log L_i(\theta)}{\partial \theta}\right) \quad (5)$$

θ_{t+1} = estimation score of ability in existing round.

θ_t = estimation score of ability in previous round.

$I(\theta)$= ability information function.

$\left(\frac{\partial \log L_i(\theta)}{\partial \theta}\right)$ = a respondent's estimation score of *maximum likelihood*.

Calculation is done until no changes may appear in the previous and last rounds or convergence. Criteria of such convergence is 0.05 or 0.01, even can be lower to 0.001. By having convergent calculation, estimation score of ability (θ) can be found or determined.

Method of ability estimation of EAP

Lord (1986) described probability of using Bayes estimation because of a tradition in education to assess similar groups of test takers using parallel or same tests from year to year. In this case we can make good description about ability frequency distribution for further groups of test takers. We can also do ability estimation of using Bayes approach through estimating procedure of Bayesian hierarchy. Such procedure, however, is difficult to implement because of lack of computer program available for it. "Researchers have adopted more pragmatic approaches in which the Bayes approach is seen as a tool to improve parameter estimation" (Baker and Kim, 2004).

In line with the latest improvement of computing system which is simpler but more sophisticated, an estimation does not use integral any longer but is based on discrete distribution. Even estimation done using EAP method could predict level of ability for all correct or wrong responses. EAP is a method applied through average calculation of *posterior* distribution. EAP calculation is done through *"Mislevy Histogram"*; its description does not show area width in a curve (Baker and Kim, 2004).

1. Determine score θ which is specifically called as *quadrature nodes* as shown in abscissa.
2. In ordinate it shows density. Usually such density and quality is taken from fixed normal distribution.
3. In "Mislevy histogram" it is assumed as normal distribution so that score or value X_k can be identified, value $A(X_k)$ shows distance between X_k and another X_{k+1}. If X_k determines same distance, value or score of $A(X_k)$ can be identified by: one divided by number of nodes. However, if the distance of X_k is not the same as $A(X_k)$, the value of $A(X_k)$ is the deviation between X_k and X_{k+1}.
4. Calculating score of L (X_k)

X_k is the same as θ_k, L is *likelihood* function of participants' ability, formula in for of multiplication as shown in formula 2.

$$L(X_k) = \prod_{i=1}^{n} P_i(X_k)^{u_{ij}} Q_i(X_k)^{1-u_{ij}} \quad (6)$$

$L(X_k)$=each participant's score of maximum ability.

X_k = θ_k= level of ability gained from formula 1
P_i = probability of correct answer
Q_i = probability of wrong answer

(v) Calculating score of ability estimation

$$E(\theta_j|U_j,\xi) = \hat{\theta}_j = \frac{\sum_{k=1}^{q} X_k L(X_k) A(X_k)}{\sum_{k=1}^{q} L(X_k) A(X_k)} \quad (7)$$

$E(\theta_j|U_j,\xi) = \hat{\theta}_j$is the average level of ability with identified requirement that the test takers' responses in the scoring is 0 or 1.

$E(\theta_j)$ = score of ability expected

ξ = score of item parameter

q = number of node (quadrature point).

Method of maximum likelihood estimation and method of expected A posteriori

The difference with both methods lays on the ability to estimate test takers' ability. *Maximum Likelihood Estimation* is not able to analyze test takers' ability if all answers they make are correct or wrong and estimating process is done with *iteration*. In contrast, *Expected A Posteriori* is able to calculate test takers' ability although they have all wrong or correct answers in such a test. In addition, the calculating process is done without *iteration* which is based on average scores of answers made by each test taker after answering certain number of items. Baker and Kim (2004) explain that *Expected A Posteriori* method uses prior data made by a particular program as "artificial data" using formula:

$$\bar{f}_{ik} = \sum_{j=1}^{N} \left[\frac{L(X_k)A(X_k)}{\sum_{k=1}^{q} L(X_k)A(X_k)} \right] \quad (8)$$

$\bar{f}_{ik}$= artificial data "*artificial examinee*" for each participant's ability (X_k)

X_k = θ_k= level of ability

$A(X_k)$ = quality, distance X_k with X_{k+1}

q = number of nodes (quadrature point) referred to level of ability.

$$\bar{r}_{ik} = \sum_{j=1}^{N} \left[\frac{u_{ij}L(X_k)A(X_k)}{\sum_{k=1}^{q} L(X_k)A(X_k)} \right] \quad (9)$$

$\bar{r}_{ik}$= artificial data "*artificial item*" correct answer in item -i in participant's level of ability (X_k)

X_k = θ_k= level of ability

$A(X_k)$ = quality, distance X_k with X_{k+1}

q = number of nodes (quadrature point), referred to level of ability.

To achieve research objectives, a numerical thinking test for students of Senior High School in Lombok-NTB was standardized applying item-response theory of two-parameter logistic model. Research was done at twenty one state senior high schools in Lombok, West Nusa Tenggara (NTB) Province with tenth year students taken as samples. The schools were situated in four regencies/cities; Mataram City, West Lombok Regency, Central Lombok Regency and East Lombok Regency. Data of this research were collected in 2008. This was an experimental research by analyzing data with BILOG MG software ver 3.0.

Independent variables were used in this research, employing ability estimation methods; *Maximum Likelihood Estimation* (MLE) and *Expected A Posteriori* (EAP). Both of them are different in method of calculating ability estimation.

Moderator variable can strengthen or weaken independent variable, referred to as second independent variable. In this research, moderator variable had many items in test packages, grouped into two test packages consisting 25 and 50 items each. Dependent variable in this research was the accuracy of estimation, limited on ability estimation variance with a unit of measurement of ability estimation result (θ).

Variable independent was grouped into two; method of ability estimation (θ) of MLE and method of ability estimation (ϑ) of EAP. Moderator variable; number of items in a test package was divided into the one with 25 items and the other with 50 items. Number of item which is less than 25 may bear an inaccurate result of item analysis program.

Populations of this research were respondents and test item population. Respondent populations were students, who gained data about their numerical thinking talent and test item population. Student populations consisted of 21 ninth year students of senior high schools in Lombok-NTB, situated at four regencies/cities. Test item populations had 120 items and grouped into three packages; A, B and C covered 40 items each. Each of the three packages had equal number of anchor items as many as nine. Thus, number of item populations were (3 x 40) – (2 x 9) = 102 items. Number of minimum samples required in a particular analysis also depends on sort of program analysis used. There is such a program requiring at least 25 items with 500 respondents. Another one may need not less than 1000 respondents (DeMars, 2010; Naga, 1992). Furtherly, quality of test used in a research data gathering was elaborated:

Test reliability

In the phase of trial I Alpha reliability coefficient of test packages A = 0.865, B = 0.906 and C = 0.933 and in trial II = 0.657. Sugiyono elaborates criteria of correlation interpretation as 0.60-0.799 (strong) and 0.80-1.00 (very strong). Calculation results gained by either modern or classical theory have high correlation coefficient, bigger than minimum requirements in a test standardizing oriented to cognitive, namely 0.85 (Sugiyono, 2010).

Test validity

Examining quality of numerical thinking talent test by internal and external validity calculation

External validity in this research used criteria variable of Differential Aptitude Tests (DAT) test result, subtest of numerical thinking talent. In subtest of numerical DAT, all items consisted of application of arithmetic operation; while standardized test consisted of that covering arithmetic operation and deliberation. Therefore, external validity examination was done prior to item separation based on two dimensions revealed by the test. Because of this, each respondent had two score dimensions; score dimensions 1 and 2. Each of this dimension correlated with scores obtained from subtest DAT through calculation of *Pearson* Correlation. Results of tests package A, B and C consecutively show as 0.415, 0.578 and 0.421 at the same dimension. At different dimension, obtaining correlation coefficient was 0.351, 515, and 0.286. Referring to this condition, tests used in data gathering of this research had good validity. Result of test examining or calibrating according to item response theory obtained from the three test packages shows:

1. In the test package A there were dominant factors based on eigen value of 7.767 with variance of 19.181%. At the next second factor was 2.001 with variance 5.004% and third factor was 1.411 with variance 3.528%.
2. Test package B has the biggest Eigen value of 7.261 with variance 19.152%, the next second factor was 2.157 with variance 5.392% and the third one was 1.535 with variance 3.839%.
3. The biggest eigen value of test package C was 7.707 with total variance 19.268%. The next second factor was 2.021 with variance 5.054% and the third one was 1.329 with variance 3.323%.

Local independency examination was aimed at recognizing whether or not an item and another in a subpopulation of certain participants' ability characters had independency statistically. Local independency indicated by score of covariance null. In this research local independency examination was done through;

1. Examining covariance score in its matrix between theta score from 10 of theta score interval and criteria reference of small covariance score or nearly reaching null. From the result of examination through covariance calculation, upper subsample (10th interval) and lower subsample (1st interval) were obtained. It had covariance score which was high enough or not nearly reaching null: test packages A = 0.23019, B= 0.2176 and C=0.2610. Because of this, an examining process can be continued to correlation examination among items; and
2. Examining correlation matrix among items using statistical package for social sciences (SPSS), gaining result of pairs of items were correlated in the three test packages. In test package A, there were three pairs of items correlated, four pairs of items for test package B and a fair of items for test package C.

Compatibility examination was conducted to know whether or not empirical data of each test item of numerical thinking talent was compatible with two-parameter logistic model. In the second phase of calculation or calibrating process, results of test item estimation were achieved covering level of difficulty (b), test item discrimination index (a) and calculation result of *chi-Square* (χ^2) together with probability index for each item. Level of significance applied in the examination was 0.01, meaning that the test items were compatible to two-parameter logistic model with index of probability ≥ 0.01, thus interpreting that test items which are good quality are 73 items. All 73 items then were considered as "pool item", and this study ultimately applied matrix sampling. Results of model compatibility examination and local independency analysis were suitably done by Jumailiyah (2015).

In line with the objective of this study, the study aims to know the distinction based on estimating methods and estimation variance referring to moderator variable consisting of 50 and 25 items. Seventy three qualified items were taken as populations as references to having research samples. Research samples were chosen randomly in two phases; the first phase was for choosing 50 items and the second one was for the rest of 25 items.

RESULTS AND DISCUSSION

Data descriptions of test and theta (θ) score of numerical thinking talent

This research employed three packages of tests. Thus three matrixes of respondents' scores were obtained. The three packages of tests consisted of multiple choice tests containing 40 items which were scored dichotomy. The highest variance gained from test package C could reach 63.769, while central tendency used calculation average. The highest calculation average reached 21.32 from test package A and the lowest was 18.76 coming from test package C.

In this section, data about result of ability estimation (θ) prior to requirement examination of items-response theory will be delivered. Data were analyzed using estimation calculation through ability estimation method of *maximum likelihood* of two-parameter logistic model. Result of ability estimation (θ) shows that the lowest minimum score was -3.6129 gaining from test package A, while test package C could only reach -3.1351. The highest theta was 3.8667 coming from test package C; whereas, the lowest one was 3.4994 obtaining from test package B.

The highest variance was found in test package C reaching 1.230 with deviation standard 1.109, while the highest calculation average of test package C was 0.0647. Data calculation of test package C reached the lowest one but had the highest variance. Distribution of *skewness* indicated positive index in the three test packages. The three types of data tended to distribute sticking outward to the right showing that the ability of most respondents were under that belonged to average. Either the raw score distribution or the theta data gained from estimation indicated same distribution in the three test packages; more data were under calculation average.

Curtosis score was found positive, indicating that it has acute distribution vertically. Result of estimation shows that data would be pulled out or moved to center of distribution. This also happened to the above of it. Result of ability estimation calculated using *maximum likelihood* method indicated that the extremely high score would have high frequency. Theoretically, when the score was further from center of distribution, it would be getting smaller or lower, yet empirical data of this research indicate that there was a tendency of getting frequency declined.

Data descriptions of ability estimation results

Data about ability estimation results refer to estimation method of *maximum likelihood* and method of *expected a posteriori,* and data based on number of items of 50 and 25 were analyzed to see estimation variance differences. Descriptions of such data are shown in the following Table 1.

An important point can be extracted from Table 1 that there were significant differences on variable magnification data which resulted from estimation in four groups using *maximum likelihood* and *expected a posteriori.* This happened either in tests with 25 or 50 items as shown in Figures 1 and 2.

Figure 1 shows results of MLE with 50 test items on the top left indicated higher ability was bigger having higher frequency than that on the top right of distribution or estimation score was getting low as seen from -2.00 up to -3.00 which had very little frequency. On the top left +2.00 up to +3.00, it indicates high frequency. If it is compared to Figure 2, releasing estimation result of EAP method, it shows that abscissa to the left and right was unbalanced. An abscissa to the left could reach up to

Table 1. Descriptions of MLE and EAP data viewed from number of item of 50 and 25.

Statistics	Estimation methods and number of items			
	MLE 50	EAP 50	MLE 25	EAP 25
N Respondent	3159	3159	3159	3159
Range	7.9239	3.6701	7.4947	4.0719
Minimum	-4.0107	-1.0804	-3.7882	-1.4087
Maximum	3.9132	2.5897	3.7065	2.6632
SD	0.9858	0.7788	1.0771	0.9246
Variance	0.972	0.606	1.160	0.855
Skew	0.600	1.039	0.678	0.779
Curtosis	2.827	1.293	1.190	0.095

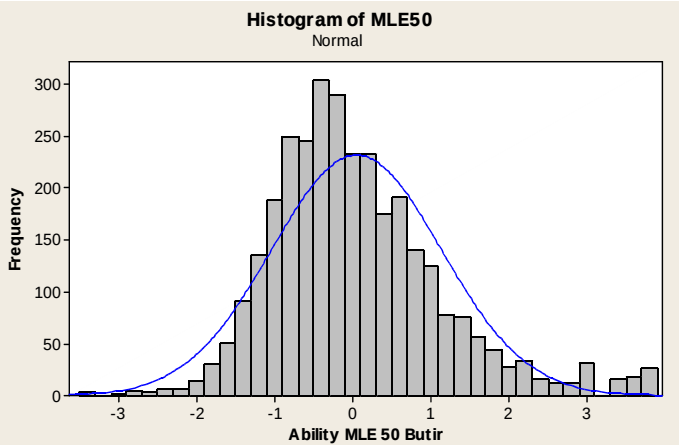

Figure 1. Ability estimation result of MLE 50 items.

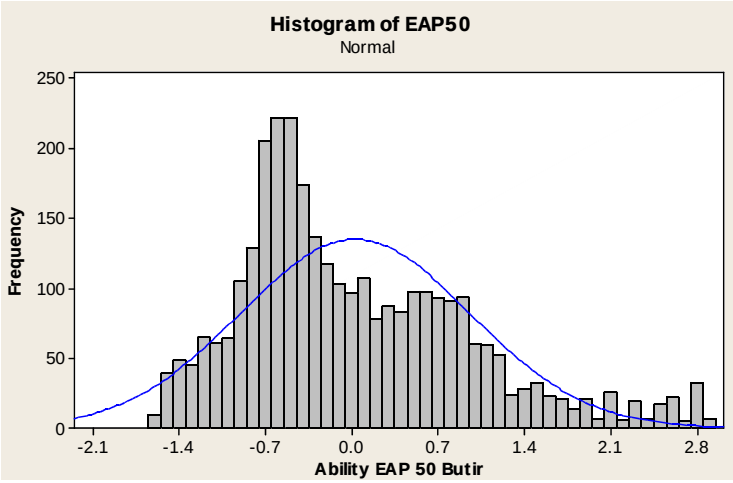

Figure 2. Ability estimation result of EAP 50 items.

+2.8, while to the right was just reaching up to -2.1. Figure 1 and 2 shows different data range of which further analysis will confirm that both MLE methods differ with EAP in variances, data range in EAP is narrower than that of MLE leading to differences in their variance.

Figures 3 and 4 describes 25 test items. Result of ability

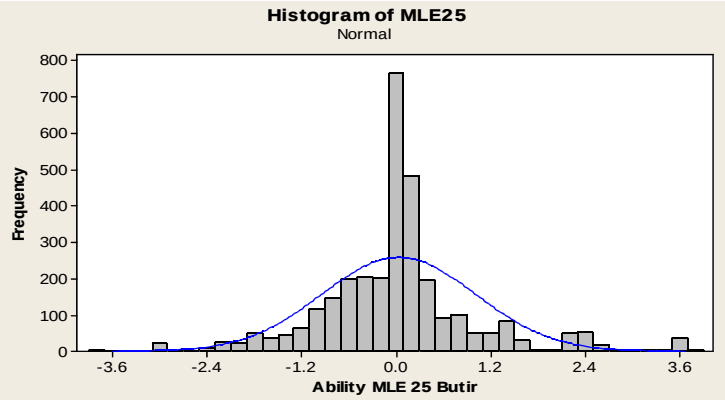

Figure 3. Ability estimation result of MLE 25 items.

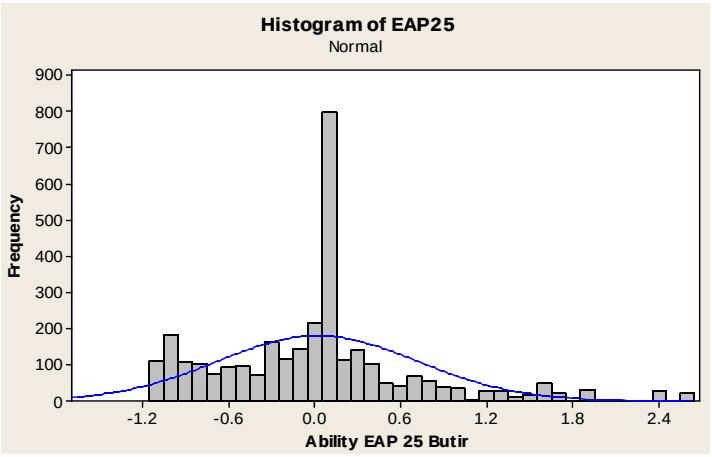

Figure 4. Ability estimation result of EAP 25 items.

estimation applying MLE as shown in Figure 3 justify that both top left and top right had high enough frequency with level of ability (θ) +3.6 and -3.6. Symmetrical index of range of estimation result indicated that it was equal either to the left or to the right. Figure 4, moreover, displays estimation result obtained through ability estimation method (θ) of *expected a posteriori* ranging from -1.2 to + 2.4 which shows different range in the graphs resulted from both methods in which MLE range is wider that EAP. Such differences will result in variance differences of both methods as it will be proven in hypothesis test. Further data exposure after estimation calculation through Bilog MG and Excel programs is presented in Table 2.

This research examined variance difference in two groups. The first group was constructed based on independent variable with methods of *Maximum Likelihood Estimation* and *Expected A Posteriori*; whereas, the second one was made up based on moderator variable consisting of 50 and 25 test items (Table 3).

Ability estimation variance ($\hat{\theta}$) of MLE method was bigger than that of ($\hat{\theta}$) gained through EAP method in each test consisting of 25 items. The value of = 0.606 and the value of = 0.972. The value of F_{hit} = 1.6023 while F_0 = 1.000. Smaller test items of 25 items each show small variances in both methods with different EAP. Then with bigger test items of 50 items, the value of = 0.806 and the value of = 1.073. The value of F_{hit} = 1.332 while F_0 = 1.000. Bigger test items of 50 items each show variances in both different methods with smaller variances in EAP method. Thus, variance differences of EAP method is smaller than MLE method in bigger or smaller items. High variance indicated measurement uncertainty. Variance can also be indicated by score distribution. The bigger score range distribution the test

Table 2. Summary of Ability estimation score ($\hat{\theta}$) variance of MLE and EAP methods viewed from number of items.

Number of items	Methods of ability estimation ($\hat{\theta}$)					
	Maximum likelihood estimation			Expected A posteriori		
	N	$\sum_{i=1}^{N}(\theta_i - \hat{\theta})^2$	Variance	N	$\sum_{i=1}^{N}(\theta_i - \hat{\theta})^2$	Variance
50 items	3159	3390.33	1.0732	3159	2545.47	0.8058
25 items	3159	3069.06	0.9715	3159	1915.29	0.6063

Table 3. Summary of statistic-based hypothesis examination result.

Variance	N	Scores of variance	Formula F	Score F	Decisions
σ^2_{MLE25}	3159	0.9715	$F = \frac{\sigma^2_{MLE25}}{\sigma^2_{EAP25}}$	1.6023	Rejected H_0
σ^2_{EAP25}	3159	0.6063			
σ^2_{MLE50}	3159	1.0732	$F = \frac{\sigma^2_{MLE50}}{\sigma^2_{EAP50}}$	1.3318	Rejected H_0
σ^2_{EAP50}	3159	0.8058			
-	-	-	-	1.1047	Rejected H_0
			-	1.3290	Rejected H_0

had, the more variances it might bear from it, yet this means that the measurement accuracy was low. This might also occur vice versa. When small variance occurs, the accuracy of estimation gets higher.

Ability estimation variance ($\hat{\theta}$) consisting of 50 test items had more variances than that ($\hat{\theta}$) consisting of 25 items obtained through *Maximum Likelihood Estimation* (MLE) method. The value of = 0.606 and the value of = 0.972 The of F_{hit} = 1.6023 while F_0 = 1.000. Smaller test items of 25 items each show small variance in both different EAP methods. Then the bigger test items or test with 50 items come with the value of 0 = 0.806 and the value of = 1.073 The value of F_{hit} = 1.332 while F_0 = 1.000. The bigger test items of 50 items each show small variance in both different EAP methods. Consequently, the test consisting less number of items may have high level of accuracy when estimation is done through MLE.

Furthermore, determining the estimation of RMS or RMSE (Root Mean Squared Error) (du Toit, 2003), there are 73 items identified as good items based on classical and modern test, randomly taken from the tests of 10, 15, 20 and 25 items in each sets (A, B, and C). These 73 items were, therefore, analyzed in both MLE and EAP.

According to Table 4, RMS or RMSE resulting from MLE and EAP implemented paired T-test. The pair of RMSE correlates 0.837 in 0.001 significance. Hypothetical test of RMSE differences between both MLE and EAP was analyzed by SPSS ver.18. The result was T = 3.060 (0.011 significance), meaning that RMS of EAP is lower than that of MLE. Other view in this study includes the number of test items to analyze, test with 50 items and test with 25 items. Both of these item groups are differentiated by MLE and EAP methods. Therefore, both methods do not show differences since both methods come with bigger variance on bigger items compared to that of test with smaller test items. A measurement appears to be good if it results in small variance. In classic theory, the more the test items the better their reliability and accuracy, thus while this study is in contrary with classical approach, it supports item response theory saying “smaller items do not mean it will result in accurate test (Embretson & Reise, 2000). Moreover, based on the accuracy of estimate of RMS, this study resulted that RMS of EAP is lower than that of MLE, and EAP resulted the estimate more accurate than that of MLE.

DISCUSSION

The study was preceded with development of numerical

Table 4. RMS estimate in both MLE and EAP, based on the number of items.

Set	Number of items	RMS	
		MLE	EAP
A	10	0.7941	0.6374
	15	0.5579	0.5373
	20	0.5057	0.4927
	25	0.4505	0.4357
B	10	0.7950	0.6125
	15	0.6429	0.5901
	20	0.5126	0.5775
	25	0.4577	0.4268
C	10	0.7418	0.5601
	15	0.6257	0.4972
	20	0.5367	0.4415
	25	0.4285	0.4071
Mean		0.5874	0.5180

reasoning talent test to show student's capability in arithmetic operation, arithmetic reasoning, basic of mathematics and its implementation in daily life. In the test development, 3159 samples and 102 items were divided into three sets of test. The quality test on the test items was based on classical test and parameter IRT 2 yielding in 73 test items that serve as “test bank”.

The items were then selected into 25 items group and 50 items group provided that they did not show “mutual exclusive” overlapping and were randomly based on the test construction. The two test item groups were tested with two methods that is, MLE of Maximum Likelihood group and EAP of Bayes group. The measurement used to show the accuracy estimation is variance from which the variance differences were tested in F significance differences while other researches presented such estimation in picture.

In some literature and studies, the two MLE estimation methods were performed through iteration that they failed to obtain maximum final value or were not convergent. The second group of Bayes employed prior distribution for their working principle. De Ayala (2009: 71) is presented in histogram, in which the range will show continuum variable such as from -4 to 4 when the number of the bar increases. This can be equalized with normal curve concept. Thus prior distribution will be the same with normal distribution.

This study uses BILOG MG ver. 3 software with its manual to guide the user, and theoretically both estimation methods were explained by Kim and Baker (2004). The parameter estimation is presented in GIBBS sample. Software that is developed to suit computer technology development will facilitate researchers in the analysis. For parameter estimation, older programs failed to show the result but newer software has been developed and improved to meet researcher's need.

This also apples to other studies using data generated by computer programs or simulation data. This study employs real sample that is, data collected from samples set forth in the study design. Study reports and journals presented data simulation obtained from SAS, and R program. Thus, software will continue to develop and facilitate researchers to obtain accurate parameter estimation.

Previous study (Borgatto et al., 2015) conducted on associating estimation method accuracy with items difficulty level found that the test item resulted in high accuracy estimation in line with classical method that is, the more the items the more accurate the result is. On the contrary, estimation in smaller items shows lower level of variance that it can be said that this study supports Embretson (2000) opinion. Study by Chen et al. (1998) concentrated on the accuracy of MLE and EAP estimation method in the implementation of Computer Adaptive Test. EAP of Bayes estimation group by varying many quadrature points came with 10, 20, 40 and 80 in prior distribution. The findings show that RMSE in MLE is smaller than the quadrature point of 10, while prior with quadrature point of 20, 40, and 80 shows relatively the same RMSE. In other word, RMSE will stay stable in quadrature point of 20. Bayes estimation made its way to a discussion that during its early implementation from Lord (1986) was considered a difficult method to understand by social science researchers. Smithson

explained the development of Bayes analysis with Markov Chain Monte Carlo (NCMC), yet it developed well particularly in social science, rare social studies employing Bayes method yet currently there are six (6) books in the phase of introduction. These books are considered containing certain type of statistics and mathematics (Smithson, 2010).

Conclusion

The following was made:

1. Variance of ability estimation () of *maximum likelihood* method is higher than that () of *expected a posteriori* one obtained from a test package consisting 25 items.
2. Variance of ability estimation () of *maximum likelihood* method is higher than that () of *expected a posteriori* one gained from a test package consisting 50 items.
3. Variance of ability estimation () of a test package consisting 50 items is higher than that of a test package having 25 items when it was done through MLE method.
4. Variance of ability estimation () of a test package consisting 50 items is higher than that of a test package having 25 items when it was done through EAP method.
5. RMS of EAP method is lower than that of MLE method in test item 10, 15, 20, and 25. Findings of this research correlate to important aspects of education mainly to those numerical thinking talent researchers and test developers who apply items-response theory of two-parameter logistic model. Accuracy of estimation result which specifically focuses on estimation variance resulted from applying methods of ability estimations of *maximum likelihood* and *expected a posteriori* is supposed to furtherly concern.

The conclusion of this research is not restricted to sample collection venue or talent test used as data collecting instrument, yet it may be generally applicable to learning materials. Consequently, such findings can be used by any teachers in constructing or making their own tests. They can construct such tests with ideal quality- assured or calibrated items. For multiple choice test with five alternative choices, they can construct it in 25 items.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

Baker FB (2004). Item response theory : parameter estimation techniques. New York: Marcel Dekker.

Baker FB, Kim SH (2004). Item response theory: Parameter estimation techniques. CRC Press.

Borgatto AF, Azevedo C, Pinheiro A, Andrade D (2015). Comparison of ability estimation methods using IRT for tests with different degrees of difficulty. Commun. Stat. Simul. Comput. 44(2):474-488.

Chen SK, Hou L, Dodd BG (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. Educ. Psychol. Meas. 58(4):569-595.

De Ayala RJ, Schafer WD, Sava-Bolesta M (1995). An investigation of the Standard Errors of Expected A Posteriori Ability Estimates. Br. J. Math. Stat. Psychol. 48(2):385-405.

De Ayala R (2009). The theory and practice of item response theory. New York: The Guilford Press.

DeMars C (2010). Item response theory. Oxford: Oxford University Press.

Du Toit M (2003). IRT from SSI: Bilog-MG, multilog, parscale, testfact. Lincolnwood, IL: Scientific Software International.

Embretson SE, Reise SP (2000). Item response theory for psychologists multivariate. Mahwah, N.J.: Lawrence Erlbaum Associates Publishers.

Hambleton RK, Swaminathan H, Rogers HJ (1991). Fundamentals of item response theory. Newbury Park, Calif: Sage Publications.

Jumailiyah (2015). Pengembangan Tes Bakat Berpikir Numerikal Model Logistik Dua Parameter. Paper presented at the Seminar dan Workshop Internasional Konseling Malindo ke-4, Denpasar, Bali.

Lord FM (1986). Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory. J. Educ. Meas. 23(2):157-162.

Magno C (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. Int. J. Educ. Psychol. Assess. 1(1):1-11.

Makridakis S, Wheelwright SC, McGee VE (1999). Metode dan aplikasi peramalan (Method and Application of Estimation). Jakarta: Binarupa Aksara.

Naga DS (1992). Pengantar teori sekor pada pengukuran pendidikan (introduction to scoring theory in educational measurement). Jakarta: Gunadarma.

Reise SP, Revicki DA (2014). Handbook of item response theory modeling: Applications to typical performance assessment. New York: Routledge.

Smithson M (2010). A Review of Six Introductory Texts on Bayesian Methods. J Educ Behav Stat. 35(3):371-374.

Sugiyono D (2010). Metode penelitian kuantitatif kualitatif dan R&D (Research method of Qualitative Quantitative and R&D). Jakarta: Penerbit Alfabeta.

Swaminathan H, Hambleton RK, Sireci SG, Xing D, Rizavi SM (2003). Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates. Appl. Psychol. Meas. 27(1):27-51.