Jurnal Pendidikan Matematika RAFA
p-ISSN :2460-8718
e-ISSN : 2460-8726

Available online at:
http://jurnal.radenfatah.ac.id/index.php/jpmrafa
June 2025, 11(1): 42-57

Estimating domain score using bayesian method and bayesian modal viewed from many domain item samples

Michrun Nisa Ramli[1)], Muslimahayati[2)]

[1),2)] Mathematics Education Study Program, UIN Sulthan Thaha Saifuddin Jambi, Jambi, Indonesia
email: [1)]michrunnisa@uinjambi.ac.id, [2)]muslimahayati@uinjambi.ac.id
(Received 30-10-2024 , Reviewed 04-06-2025 , Accepted 15-06-2025)

Abstract

The purpose of this research was to compare two methods of domain scores estimation namely Bayesian method and Bayesian Modal method were derived from many sample items taken from the domain. The research was a comparative quantitative studies of the junior high school students in Jambi. The data were analyzed using variances difference test domain scores between the two populations. The results were as follow: 1) the Bayesian was the most accurate method for domain scores estimation, 2) the more samples that were taken from the domain items, the more accurate the domain scores would be obtained. The results of this study strongly suggest that in order to determine the level of mastery for a specific material, students do not need to work on all the items but only some of them are taken from the population of items that contain this material
***Keywords** : Bayesian, Domain Scores, Modal Bayesian*

Abstrak

Tujuan penelitian ini adalah untuk membandingkan dua metode estimasi skor domain, yaitu metode Bayesian dan metode Modal Bayesian, yang berasal dari banyak item sampel yang diambil dari domain tersebut. Penelitian ini merupakan studi kuantitatif komparatif terhadap siswa SMP di Jambi. Data dianalisis menggunakan uji perbedaan varians skor domain antara kedua populasi. Hasilnya adalah sebagai berikut: 1) Bayesian merupakan metode yang paling akurat untuk estimasi skor domain, 2) semakin banyak sampel yang diambil dari item domain, semakin akurat skor domain yang diperoleh. Hasil penelitian ini menunjukkan bahwa untuk menentukan tingkat penguasaan materi tertentu, siswa tidak perlu mengerjakan semua item, tetapi hanya beberapa item yang diambil dari populasi item yang memuat materi tersebut.

***Kata Kunci**: Bayesian, Skor Domain, Modal Bayesian*

INTRODUCTION

One effort to improve the quality of education is to improve the assessment system applied and knowing whether a learning system is running or not (Andayani & Madani, 2023; Mardapi, 2007; Musarwan & Warsah, 2022; Prastiwi et al., 2023; Rahman et al., 1993). Assessment requires good quality data so it needs to be supported by a good measurement process. Measurement will produce a score which is then interpreted into a value. High and low values are usually associated with a reference. The general reference assessment used is Norm Referred Measures and Criterion Referred Measures. In this study, the author only discusses Criterion Referenced Measures. The criteria for assessing criteria are not intended to compare one student with another, the goal is to determine the level of mastery of a subject matter by comparing it with an existing mastery criterion. Reference assessment criteria have been applied in schools.

Domains are developed from reference criterion assessments, domains are a collection of well-described mastery tests of a task that are intended to describe student status, domains can consist of several sets of items that are carefully arranged by experts, in an ideal situation if the domain has a domain definition consisting of skills and abilities that reflect mastery of a particular content area (Kern, 2007; Popham, 1974). The population item measuring instrument is a measuring instrument that contains all items that can be arranged and outside the item population there are no other items that can be used, while the sample measuring instrument contains some of the items contained in the population item measuring instrument (Naga, 2012). The interpretation of the criteria assessment will lose its validity if the domain that is expected to be able to draw conclusions about student status is not well defined, or if the sample of items taken from the domain is not representative (Nitko, 2001). From several expert opinions regarding the definition of domain, a synthesis can be made regarding the domain, namely a collection of questions that contain all the skills and knowledge that must be possessed to master a certain content that is arranged based on the format for compiling the questions.

To be able to determine students' mastery of a particular content area, a score called a domain score is needed. Domain scores indicate students' performance in a group of domain items that represent the skills and knowledge needed to master a content area, domain scores are the percentage of correct answers if all items in the domain are given, this score can be estimated by providing test items taken randomly from the domain (Pommerich & Nicewander, 1998; Popham, 1974). From the opinions of these experts, it can be concluded that domain scores are estimated scores given to test takers to find out

whether they have mastered a domain or not with test takers only responding to representative samples of items taken from the domain, while the scores obtained are valid if all items in the domain are given. Because domain scores can only be estimated, a method for estimating this domain score is needed. Currently, there is a modern measurement theory known as Item Response Theory, this theory overcomes the weaknesses of classical measurement theory. Classical measurement and modern measurement have different characteristics, classical measurement of the characteristics of test items cannot be separated from test takers, while modern measurement of the characteristics of test items will remain the same even though the test takers are different (Naga, 2012).

Characteristics of items such as the level of difficulty of items, distinguishing power, and opportunity factors are known as grain parameters, while the characteristics of test participants are known as the ability parameters. Both of these parameters can be estimated using several methods, including the Bayesian estimation method (Expected a posterior) and the Bayesian Capital Estimation Method (maximum a posterior).

In this study two kinds of estimated words were discussed, the first was the estimated ability parameter, the second was the estimated domain score. To avoid mistakes, a different symbol is given to estimate the ability parameters and estimated domain scores. The capability parameter estimated is symbolized as θ_j while the domain score estimated is symbolized as θ^*_j.

The Bayesian estimation method or commonly also called Expected a Posterior (EAP) is an estimated method that is able to analyze or calculate the ability of participants with all response patterns such as participants answering all true or all wrong, the calculation process is carried out by calculation without iteration but based on average value. The average answer of each participant after answering a number of items (Baker & Kim, 2004.

The Bayesian Modal Estimation Method or also commonly called Maximum A Posterior (MAP) is an estimated method that uses iteration to get the estimated capability score θj. The Bayesian Modal Method uses iteration with all the pattern of response from the test participants, namely for all correct responses or all the wrong responses that will be estimated, the Bayesian Estimation Procedure is always converging for each possible grain response pattern (Baker & Kim, 2004. This study aims to determine the difference in the variance of domain scores through the estimated domain score using the Bayesian estimation method and the Bayesian modal estimation method as well as to find out the difference in the variance of domain scores in groups of students who work on a

series of questions with different test lengths on the class IX student algebra mastery test in Jambi City.

METHODS

The domain score estimation in this study was carried out using several steps, starting by determining the respondent population and the item population, then the grain response data was used to estimate the score scores of each respondent and the item parameter using the Bayesian estimation method and the Bayesian modal estimation method. After the ability score is obtained, then a domain score estimate calculation is carried out, because this research uses the theory of response items, the assumption of the theory of response items must be fulfilled, then hypothesis testing is carried out using a variance's different test. This study uses Item Response Theory (IRT), which essentially aims to address the weaknesses of classical measurement (Sudaryono, 2011). Item response theory is a psychometric model used to analyze the relationship between individual abilities and their responses to test items (Embretson & Reise, 2013)

The population of this study is a class IX junior high school/MTs student in Jambi City and specific content that will be measured the level of mastery is the algebra material learned at the junior high school level. The population of items or item domain of items consisting of 22 items of algebra is formed into three different question devices, namely the questions arranged based on the percentage of all items in the domain. In this study three types of percentages were used, namely 40%, 60% and 80%, because the number of sample items taken from the domain was recommended at least 40% (Popham, 1974).

In the theory of item response it is known that there are three assumptions that must be met, namely invarian, unidimensionality and local independence. Invarian is a characteristic of test items that do not depend on the distribution of the parameters of the capability of the test participants, unidimensionality means that each test item only measures one local ability and independence that is no relationship between one test item and another test item (Naga, 2012; Retnawati, 2014). Item response theory needs to determine the item characteristic model used, namely one item parameter (1PL), two item parameters (2PL) and three item parameters (3PL), or other models (Lord, 1990). In the theory of item response it is necessary to determine the grain characteristic model used, in this study the logistics model of two parameters or called L2P is chosen.

The instrument used was a multiple choice algebra mastery test with 4 answer choices. The test instrument was built from the collection of algebra mastery items from the two Packages of the National Mathematics Mathematics Examination Question

Package, the response data of the two question packages were estimated by the grain parameter and test the suitability of the L2P model, the analysis was carried out using the Bilog Mg version 3.0 program. Analysis of Model Model Model on the two selected national exam questions needs to be done because the items containing algebra and in accordance with the L2P model are used for domain items. After the grains are in accordance with the selected model, then an equalization of grain parameters is carried out using the mean and sigma equivalence methods.

After the respondent or sample responds to the test tools he received, an estimated ability score of each respondent and grain parameter, namely the level of difficulty, different power and true chance factor. Estimation was made using two estimation methods, namely the Bayesian estimation method and the Bayesian Modal Estimation Method.

Table 1. Research design

Estimation Method	Test Length 40% (A1)	Test Length 60% (A2)	Test Length 80% (A3)
Bayesian (B1)	$\theta^*_1,\theta^*_2, ..., \theta^*m_1$ $S^2B_1A_1$	$\theta^*_1,\theta^*_2, ..., \theta^*m_2$ $S^2B_1A_2$	$\theta^*_1,\theta^*_2, ..., \theta^*m_3$ $S^2B_1A_3$
Bayesian Modal (B2)	$\theta^*_1,\theta^*_2, ..., \theta^*m_4$ $S^2B_2A_1$	$\theta^*_1,\theta^*_2, ..., \theta^*m_5$ $S^2B_2A_2$	$\theta^*_1,\theta^*_2, ..., \theta^*m_6$ $S^2B_2A_3$

Notes: $\theta^*_1,\theta^*_{2, ...,}\theta^*_{m1}$= respondent's domain score;

$S^2B_1A_1$ = variance of the group of respondents who took the A1 test and the Bayesian estimation method.

$S^2B_2A_1$ = variance of the group of respondents who took the A1 test and the Bayesian modal estimation method

Furthermore, the score score of all respondents that have been estimated using the Bayesian estimation method and Bayesian modal are used to estimate the domain score. The theory of response item provides the right method to estimate the domain score using the test participant response to the test device and know the grain parameter θ^*_j (Pommerich & Nicewander, 1998. Through this approach, the domain score for test participants with a item domain consisting of only multiple choice items can be estimated to use the following formula:…

$$\theta^*_j = \frac{1}{s}\sum_{s=1}^{S} P_s(\theta_j) \quad \text{... (1)}$$

Notes: θ^*_j= domain score; θ_j= the ability score of the j-th test taker; $P_s(\theta_j)$ = the probability of correctly answering the s-th item domain on ability θ_j; s = 1,2,3,…S; S = number of items in the domain.

Opportunity to answer correct $P_s(\theta_j)$ calculated using the formula for the probability of answering correctly for the two-parameter logistic model, namely:

$$P_i(\theta_j) = \frac{e^{Da(\theta_j - b_i)}}{1 + e^{Da(\theta_j - b_i)}} = \frac{1}{1 + e^{-Da(\theta_j - b_i)}} \quad \ldots(2)$$

Note: $P_i(\theta_j)$ = opportunity to answer item i correctly on ability θ_j; θ_j = the ability of participant j on each item; j = 1, 2, 3,…,N where N = number of test takers; a_i = difference power of item i; b_i = level of difficulty item i.

With A_I and B_I is a grain parameter that applies to the domain and ability parameter θ_j is the estimated test participant after responding to the test devices they receive. The domain score of each test participant is calculated using the capability score that has been estimated using the Bilog MG program with the Bayesian modal estimation method and the Bayesian estimation method.

After obtaining the estimated value of the domain score θ^*_j , then the calculation of the F value is done to test the hypothesis. Calculation of F values through the calculation of the population variance of each group (Iriawan & Astuti, 2006) with the following formula:

$$S_k^2 = \sum_{j=1}^{N_{kj}} \frac{\left(\theta^*_j - \overline{\theta^*_{kj}}\right)^2}{N_{kj}} \quad \ldots(3)$$

Notes: S_k^2 = population variance of the k-th group domain scores; θ^*_j = j participant's domain score value; $\overline{\theta^*_{kj}}$ = the mean value of the k-group domain scores; N_k = the number of respondents in the k-th group.

Then proceed with calculating the value of F with the following formula:

$$F = \frac{S_1^2}{S_2^2} \; or \; \frac{S_2^2}{S_1^2} \quad ..(4)$$

Notes . $\frac{bigger \text{ var} iant}{smaller \text{ var} iant}$

RESULT AND DISCUSSION

The application of the theoretical response items must be preceded by proof that the assumption of the theoretical response items has been fulfilled. The assumption of the theory response items is unidimensional, local independence and invarians. To prove unidimensional assumptions, factor analysis is carried out. Because the unidimensional test wants to find out whether each test item only measures one ability, the method used is the analysis of the main components. One way to find out how many factors, this analysis can be done, among others, based on the Eigen value and based on the Scree Plot. Eigen value is the total variance described by each factor, a factor with an eigen value >1 can be maintained, while the Scree Plot is an eigen value plot that is associated with a number of factors (Djaali & Pudjiono, 2007. Based on the results of the factor analysis in the three sets of the questions answered by the test participants, there are 2, 4, and 6 eigen values that are greater than 1. But of the three devices about these questions, there is only one of the most dominant factors, namely the first factor, so that the dominance of the first factor is able to provide support for evidence of unidimensionality response data. Furthermore, the Eigen value was observed using a Scree Plot, it appears that the eigen value begins to slop the third factor for all the question devices used. This shows that there is only 1 dominant factor in the algebra mastery test. Thus it can be said that the items of the algebra mastery test from all the questions about the question have met unidimensional requirements.

In the response theory, each item is independent, thus there is no relationship between one point and another so that the chance of answering correctly in one point is not influenced by the opportunity to answer correctly other items, as well as the ability of one test participant to another test participant. The ability to answer correctly a test participant does not depend on the ability to answer the correct test participants. The assumption of local independence can be proven by calculating the covariance value between the subpopulation of the score of the test participants, because the meaning of the unidimensional definition must be based on the assumption of local independence and the test item device will be unidimensional if the test participant with the same ability has a covariance value between items in the device is zero (Hambleton et al., 1991; Hambleton & Swaminathan, 1991).

The assumption of local independence is proven by observing the covariance value between test participants' ability scores. The test participants' ability scores or those obtained from the responses of each set of questions are sorted from the smallest to the largest, then arranged into ten intervals, and then the covariance between the intervals is

calculated using SPSS software version 16. For each set of questions, the test participants' ability score intervals are named C1, C2, C3 to C10.

After calculating the covariance between the ability score intervals for question set A1, question set A2, and question set A3, all are close to zero. So it can be said that the three sets of questions have met the requirements of local independence. The next assumption of Item Response Theory is invariance, invariance means that the characteristics of the question items do not depend on the distribution of the test participants' ability parameters and the parameters that characterize the test participants do not depend on the characteristics of the question items (Retnawati, 2014). A person's ability will not change just because they do tests with different levels of difficulty and the parameters of the question items will not change just because they are tested on groups of test participants with different levels of ability.

In the assumption of unidimensional item response theory, it has been proven that all question devices used in data collection in this study have met the unidimensional requirements, so that it has automatically been proven that all question devices have maintained the invariant properties of the Item Response Theory. The requirement of unidimensional items is intended to maintain invariance in the item response theory, if the item measures more than one dimension, then the answer to the item will be a combination of various abilities in the participants (Naga, 2012). After all the assumptions of the Item Response Theory are met, the hypothesis is tested using the variance difference test with the variance comparison analysis technique of the two groups, the domain score is estimated so that θ^*_j is obtained for each respondent.

In this study, the significance level used is $\alpha = 0.05$ and uses a one-way test.

Hypothesis: $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$

$H1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$

There are nine hypotheses in this study, namely:

First hypothesis

The variance of the domain scores of the group of students who took the 9-item test was greater than that of the group of students who took the 14-item test for the Bayesian modal domain score estimation method.

Second hypothesis

The variance of the domain scores of the group of students who took the 9-item test was greater than that of the group of students who took the 18-item test for the Bayesian modal domain score estimation method.

Third hypothesis

The variance of the domain scores of the group of students who took the 14 item test was greater than that of the group of students who took the 18 item test for the Bayesian modal domain score estimation method.

Fourth hypothesis

The variance of the domain scores of the group of students who took the 9-item test was greater than that of the group of students who took the 14-item test for the Bayesian Modal domain score estimation method.

Fifth hypothesis

The variance of the domain scores of the group of students who took the 9-item test was greater than that of the group of students who took the 18-item test for the Bayesian Modal domain score estimation method.

Sixth hypothesis

The variance of the domain scores of the group of students who took the 14-item test was greater than that of the group of students who took the 18-item test for the Bayesian Modal domain score estimation method.

Seventh hypothesis

The variance of domain scores through the Bayesian Modal method is greater than the Bayesian method for groups of students who take the 9 item test.

Eighth hypothesis

The variance of domain scores through the Bayesian Modal method was greater than the Bayesian method for the group of students who took the 14 item test.

Ninth hypothesis

The variance of domain scores through the Bayesian Modal method is greater than the Bayesian method for groups of students who take the 18-item test.

The summary of the F values and the testing of the nine hypotheses are summarized in the following table:

Table 2. Hypothesis Testing Table

Hypothesis	Many Respondents	F Nilai value	Decision
First	1081 and 1049	1.0664	Reject Ho
Second	1081 and 1068	1.10608	Reject Ho
Third	1049 and 1068	1.0371	Reject Ho
Fourth	1081 and 1049	1.1106	Reject Ho
Fifth	1081 and 1068	1.142	Reject Ho
Sixth	1049 and 1068	1.0282	Reject Ho
Seventh	1081	2.254	Reject Ho
Eighth	1049	2.1368	Reject Ho
Ninth	1068	2.1554	Reject Ho

F_count=1.0664 > F_table=1 H0 is rejected, it can be concluded that the variance of the domain score θ^*_j of the group of students who took the 9-item test is greater than the group of students who took the 14-item test for the Bayesian estimation method. And so on for all hypotheses resulting in a decision to reject the null hypothesis. So it can be concluded that the group of students who took the test with more items produced a smaller variance of domain scores, than the group of students who took the test with fewer items for the Bayesian domain score estimation method and the Bayesian modal score estimation method. And the variance of the domain score through the Bayesian Modal method is greater than the Bayesian method for all groups of students who worked on the 9-item, 14-item, and 18-item test sets.

The accuracy of an estimate can be seen from the magnitude of the variance, the greater the variance, the less accurate the estimate. If the parameter value obtained through estimation contains a large variance, it means that the parameter value is not sharp or accurate enough (Naga, 1992).

The results of the study showed that the variance of the domain score of the group of students who worked on the test with 9 questions was greater than the variance of the domain score of the group of students who worked on the test with 14 questions for the Bayesian domain score estimation method or it can be concluded that the domain score of the test with 14 questions is more accurate than the domain score of the test with 9 questions estimated using the Bayesian estimation method. This can be seen visually from Figure 1, from Figure 1 it can be explained that the length of the domain score box of the group of students who worked on 9 questions is longer than the length of the domain

score box of the group of students who worked on 14 questions, the longer the box indicates that the data is more spread out.

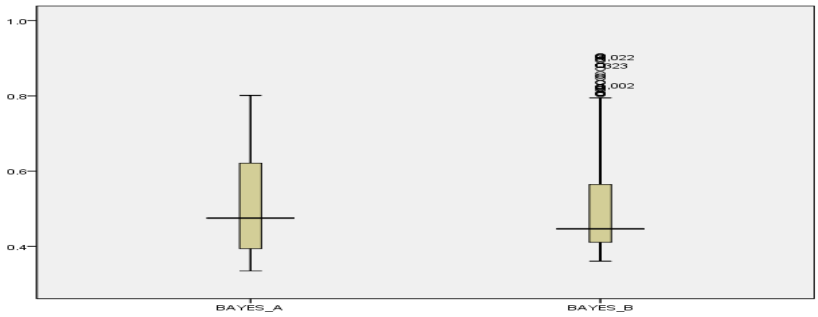

Figure 1. Boxplot domain score set A1 and set A2 using the Bayesian Method

This study also concluded that the 18-item test domain score was more accurate than the 9-item test domain score estimated using the Bayesian estimation method. Visually it can be explained by **Figure 2**, it can be seen that the length of the domain score box for the student group who worked on 9 items is longer than the domain score box for the group of students who worked on 18 items, this means that the domain score for the group of students working on 9 items is more spread out.

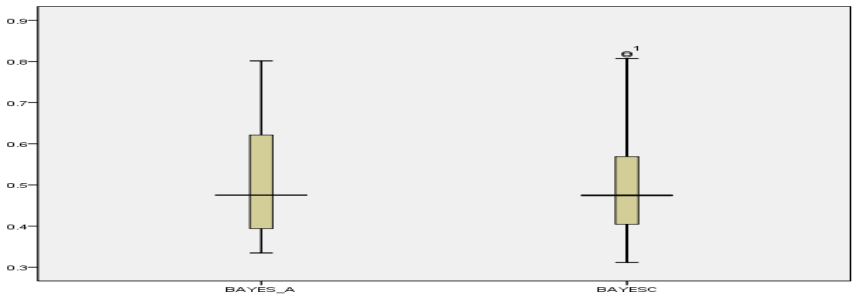

Figure 2. Boxplot domain score set A1 and set A3 using the Bayesian Method

Furthermore, the results of the study also showed that the 18-item test domain score was more accurate than the 14-item test domain score which was estimated using the Bayesian estimation method. Visually it can be explained in Figure 3, it can be seen that although the domain score box for the group of students working on 18 items is longer than the domain score box for the group of students working on the 14 item questions, the

box for the A3 question set is more symmetrical than the box for the A2 question set and does not. there is an outlier value in the domain score of the A3 question set.

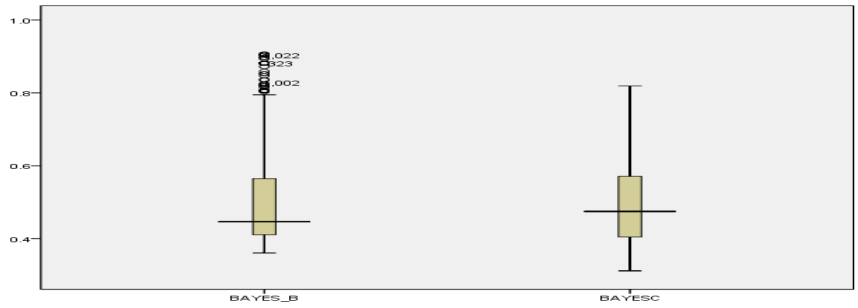

Figure 3. Boxplot domain score set A2 and set A3 using the Bayesian Method

The results also show that the 14-item test domain score is more accurate than the 9 -item test domain score estimated using the Bayesian Modal estimation method. This can be seen visually from **Figure 4**, it can be explained that the length of the domain score box for the group of students who worked on 9 items was longer than the length of the domain score box for the group of students who worked on 14 items, a longer box indicates the data is more spread out.

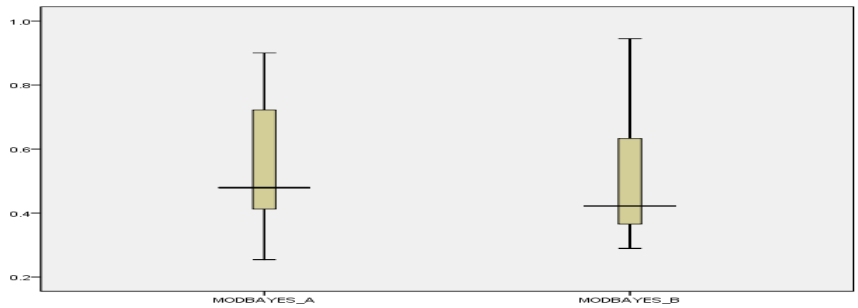

Figure 4. Boxplot domain score set A1 and set A2 using the Bayesian Modal Method

Furthermore, the results of the study also showed that the 18-item test domain score was more accurate than the 9-item test domain score which was estimated using the Bayesian Modal estimation method. This can be seen visually from **Figure 5**, it can be explained that the length of the domain score box for the group of students working on 9 items is longer than the length of the domain score box for the group of students working on 18 items, a longer box indicates the data is more spread out.

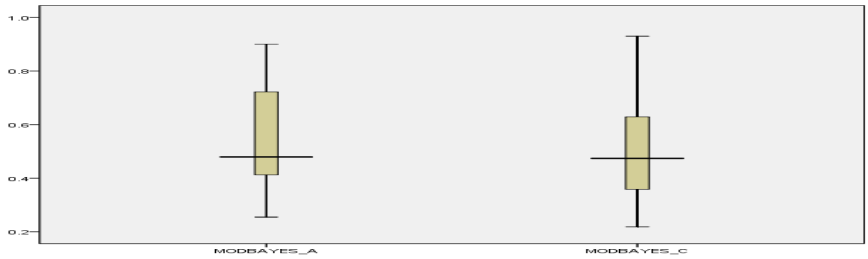

Figure 5. Boxplot domain score set A1 and set A3 using the Bayesian Modal Method

The results also show that the 18-item test domain score is more accurate than the 14-item test domain score estimated using the Bayesian Modal estimation method. This can be seen visually from **Figure 6**. It can be explained that the domain score data of the student group working on 18 items is more symmetrical than the domain score data of the student group working on 14 items, and the domain score data of the student group working on the 14 items is more spread out than the data. the domain score of the group of students who worked on 18 items when viewed from the length of the box.

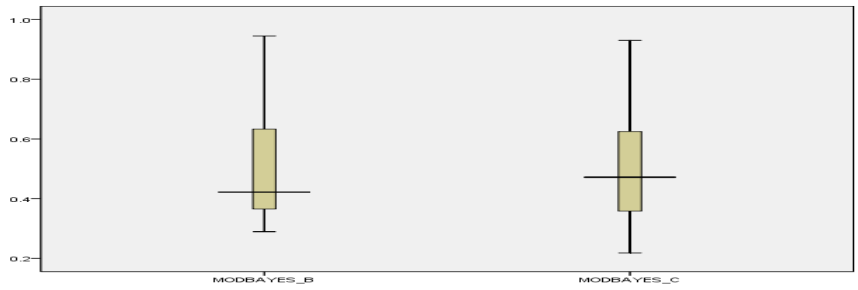

Figure 6. Boxplot domain score set A2 and set A3 using the Bayesian Modal Method

This study also found that the domain score obtained using the Bayesian estimation method was more accurate than the domain score obtained using the Bayesian Modal estimation method for the test length of 9 items, 14 items and 18 items. This can be explained visually from the Boxplot form in **Figures 7**, **8** and **9**.

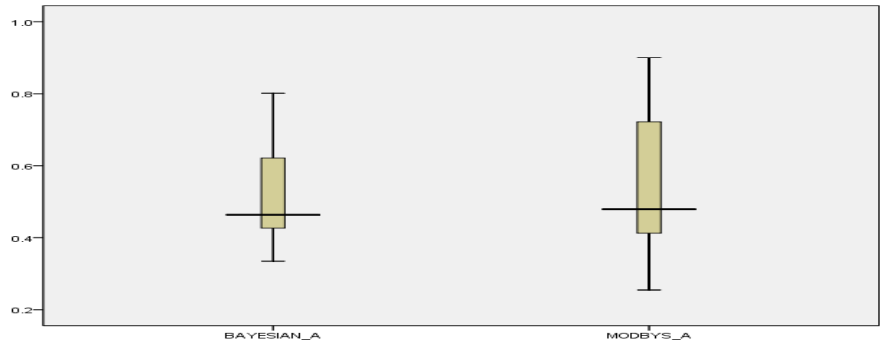

Figure 7. Boxplot domain score set A1 question using Bayesian Method and Bayesian Modal

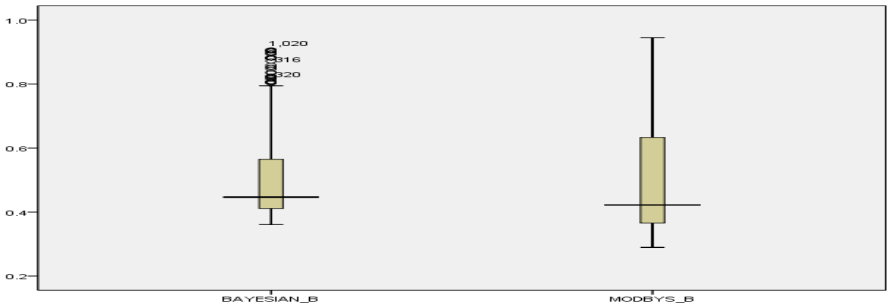

Figure 8. Boxplot domain score set A2 question using Bayesian Method and Bayesian Modal

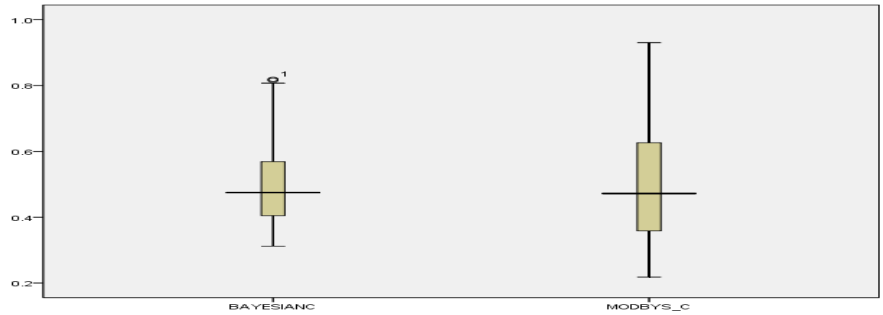

Figure 9. Boxplot domain score A3 question set using Bayesian Method and Bayesian Modal

From **Figures 7**, **8**, and **9**, it can be seen that the length of the domain score data box estimated by the Bayesian Modal method is longer than the length of the domain score data box estimated by the Bayesian method. Although the domain score data estimated by the Bayesian method has outlier values, this does not significantly affect the

resulting variance. In all problem sets, the domain score estimation by the Bayesian estimation method produces a smaller variance compared to the domain score estimation results by the Bayesian Modal method.

CONCLUSION

Based on the findings and considering the limitations of this study, it can be cconcluded as follows that longer test items produce smaller domain score variances, so it can be said that the estimation of domain scores for longer test items is more accurate than the estimation of domain scores for shorter test items, both for domain scores estimated using the Bayesian estimation method and the Bayesian modal method. This is in line with the results of (Hasanah et al., 2024) research which states that there is a relationship between ability parameters and test length. This shows that the greater the percentage of items taken from the domain, the more accurate the domain score estimation will be. Furthermore, it can also be concluded that the Bayesian estimation method is more accurate than the Bayesian Modal estimation method for all test lengths.

Many weaknesses were found in this study, for that the author would like to provide advice to researchers who are interested in researching domain scores, namely to create their own tests that meet the domain requirements in order to get more domain test items so that more domain item samples are taken. The BILOG MG software will be more sensitive if using many items so that it will be easy to test the suitability of the model.

REFERENCES

Andayani, T., & Madani, F. (2023). Peran penilaian pembelajaran dalam meningkatkan prestasi siswa di pendidikan dasar. *Jurnal Educatio FKIP UNMA*, *9*(2), 924–930. https://doi.org/10.31949/educatio.v9i2.4402

Baker, F. B., & Kim, S. H. (2004). *Item response theory parameter estimation techniques*. Marcel Dekker Inc.

Djaali, & Pudjiono. (2007). *Pengukuran dalam bidang pendidikan*. Gramedia Widia Sarana.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Hambleton, R. K., & Swaminathan, H. (1991). *Item response theory*. Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publication Inc.

Hasanah, S., Pomalingo, D. Z., Kurnia, A., Dwinata, A., Matematika, P., Maritim Raja, U., & Haji, A. (2024). Pengaruh panjang tes dan ukuran contoh pada pendugaan pa-

rameter model IRT. *SAP (Susunan Artikel Pendidikan)*, *8*(3), 408–412. https://doi.org/10.30998/sap.v8i3.19663

Iriawan, N., & Astuti, S. P. (2006). *Mengolah data statistik dengan mudah menggunakan minitab 14*. CV. Andi Offset.

Kern, H. (2007). *The estimation of domain score through IRT methods on a mathematics placement test*. Clemson University.

Lord, F. M. (1990). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Publishers.

Mardapi, D. (2007). *Teknik penyusunan instrumen tes dan nontes*. Mitra Medika.

Musarwan, M., & Warsah, I. (2022). Evaluasi pembelajaran (konsep, fungsi dan tujuan) sebuah tinjauan teoritis. *Jurnal Kajian Pendidikan Islam*, 186–199. https://doi.org/10.58561/jkpi.v1i2.35

Naga, S. D. (1992). *Teori skor pada pengukuran pendidikan*. Besbats.

Naga, S. D. (2012). *Teori skor pada pengukuran mental*. Nagarani Citrayasa.

Nitko, A. J. (2001). *Educational assessment of student*. Merrill Prentice Hall.

Pommerich, M., & Nicewander, W. A. (1998). *Estimating average domain scores* (Issues 98–5).

Popham, W. J. (1974). *Evaluation in education*. McCutrhan Publishing Corporation.

Prastiwi, Y. E. N., Arba'iyah, Amatullah Al Barru, A., & Syarif Hidayatullah, A. (2023). Penilaian dan pengukuran hasil belajar pada peserta didik berbasis analisis psikologi. *Bersatu: Jurnal Pendidikan Bhinneka Tunggal Ika*, *1*(4), 218–231.

Rahman, F., Hermina, D., Huda, N., Palangka Raya, I., Tengah, K., Antasari Banjarmasin, U., Selatan, K., & Author, C. (1993). *Urgensi penilaian pembelajaran di dunia pendidikan formal (hakikat, sistem, dan perencanaan)*. *5*(2), 145–162. https://doi.org/10.23971/mdr.v5i2.5064

Retnawati, H. (2014). *Teori respons butir dan penerapannya*. Nuha Medika.

Sudaryono, S. (2011). Implementasi teori responsi butir (Item Response Theory) pada penilaian hasil belajar akhir di sekolah. *Jurnal Pendidikan Dan Kebudayaan*, *17*(6), 719–732. https://doi.org/10.24832/jpnk.v17i6.62