Al-Ishlah: Jurnal Pendidikan
Vol.14, 1 (April, 2022), pp. 385-398
ISSN: 2087-9490 EISSN: 2597-940X, DOI: 10.35445/alishlah.v14i1.926

Estimating Item Parameters and Student Abilities: An IRT 2PL Analysis of Mathematics Examination

Jumini[1], Heri Retnawati[2]
[1] Universitas Negeri Yogyakarta, Indonesia; e-mail: jumini.2020@student.uny.ac.id
[2] Universitas Negeri Yogyakarta, Indonesia; e-mail: heri_retnawati@uny.ac.id

ARTICLE INFO

Keywords:

Item response theory;
Mathematics examination;
Two parameter logistic.

Article history:

Received 2021-08-05
Revised 2021-12-17
Accepted 2022-02-15

ABSTRACT

This research aims to describe the characteristics of math test instruments are tested on 373 of 10th-grade vocational high school students and to describe student abilities. The analysis was conducted using the Item Response Theory (IRT) approach with the 2 Parameter Logistic (2PL) model. The whole analysis process was conducted with the help of the R Program, SPSS and Excel. This finding shows that 2 items did not fit the 2 PL model. Of the 28 items that fit the 2PL model, 3 do not have a good discriminating index because they have negative values. Of 30 items that were tested, 3 items were very easy, 21 items were moderate, one item was difficult, and five items were very difficult. Student abilities ranged between -4.69 logit and 4.09 logit with an average of 0.05 logit. Based on the category, the proportion of students with moderate ability reached 63.54%, high ability 15.28%, very high ability 1.34%, while students with low ability were 19.57%, and very low ability was only 0.27%. The maximum information function of this test was 29,37 (has good category) on the ability (θ) of -0,8 logit and SEM of 0.18 logit. Overall, the test was suitable for students with abilities between -2,6 and 2,4 logit. Based on these results, 25 items can be included in the question bank with suggestions for improvement in the proportion of difficulty level to reach a normal curve balance.

Corresponding Author:
Jumini
Universitas Negeri Yogyakarta, Indonesia; e-mail: jumini.2020@student.uny.ac.id

1. INTRODUCTION

Evaluation plays an important role for the education world to provide information to assess the outcomes of students and improve the material and educational programs (Worten & Sanders, 1987). Four terms closely related to evaluation in the world of education are measurement, testing, assessment and evaluation. All of which are related to each other and form a hierarchy. Tests are conducted to determine students' abilities. Measurement plays a role in quantifying these abilities. Assessments explain and interpret measurement results, while evaluation determines the value or implications of a

policy or decision. Thus, an evaluation activity will involve measurement and assessment (Mardapi, 2012: 6-7).

Assessment of learning outcomes by educators is the process of collecting and processing information to measure the achievement of student learning outcomes in aspects of attitudes, knowledge, and skills which are conducted in a planned and systematic manner to monitor processes, learning progress, and improve learning outcomes (Direktorat Pembinaan Sekolah Menengah Kejuruan Kemdikbud RI, 2018). One component of assessing learning outcomes in schools is a formative assessment in the form of the Mid-Semester Examination conducted in the middle of the odd semester or even semester. Formative assessment is conducted throughout the semester, each item or sub-subject matter to determine the success and failure of the learning process (Istiyono, 2018). This assessment aims to improve learning strategies (Mardapi, 2012: 16).

Assessment and learning during the pandemic are conducted remotely with various methods and technological means to prevent the transmission of Covid-19. So is the Middle Semester Exam, where implementation uses a model Computerized Based Test (CBT), and students can do it online using devices such as handphones and computers from their homes. Although the CBT model has weaknesses, especially in terms of supervision, it also provides convenience to students and schools in implementing the test during the pandemic situation. This is because the CBT model can be implemented remotely, is easily accessible, and the order of questions and answer choices can be randomized to minimize the potential for fraud during implementation. Because this Mid-Semester Examination is a formative assessment, the results can be used to improve learning strategies, especially learning currently being implemented, namely distance learning.

According to Azwar (2016), a new test can be successful in running out its measuring function if it can provide accurate measurement results. Thus, the quality of the test is largely determined by the quality of the items. As with other assessment instruments, the instruments used in the Mid-Semester Examination must also meet the requirements and characteristics of a good instrument . One of them is instrument reliability which is the concept of the extent to which the measurement results can be trusted (Azwar, 2019). In objective type item analysis, item quality is also seen from several parameters, namely difficulty index, discriminant index and distractor effectiveness. The conclusion regarding the item's quality will lead to a decision regarding whether or not the item can be used, revised or replaced and whether the item meets the requirements to be included in the item bank or question bank (Azwar, 2016).

Based on the results of discussions with the teacher of mathematics SMK joined in MGMPs, obtained information that items analysis performed by the teacher during this time is merely a description of difficulty index and discriminant index of the Classical Test Theory approach. Item analysis has never been conducted, including estimating the instrument reliability and determining the characteristics of the items using the Item Response Theory (IRT)approach. According to Retnawati (2014), Classical Test Theory has several weaknesses, including group dependence, where the discriminant index and the difficulty index of the items are highly dependent on the average ability level, range and distribution of student abilities used as samples in the analysis. In addition, the student's acquisition score is highly dependent on the selection of the test used, not on the ability of the student. Some of the Classical Test Theory weaknesses were overcome by developing IRT (Mardapi, 2012).

The three main assumptions in the IRT are unidimensionality, local independence and parameter invariance (Hambleton & Swaminathan in Retnawati, 2014)). Local independence is the probability of answering correctly must be independent of each other while the unidimensional substances measured come from one dimension (Mardapi, 2012). Parameter invariance can be interpreted that a person's ability will not change because of solving test questions with different levels of difficulty, and the parameters of the test items will not change because they are tested on a group of test respondents with different levels of ability (Retnawati, 2014). The correlation between the probability of answering correctly on a scale of ability P(θ) is expressed by a correlation with item parameters that are used in which the number of parameters of this item to determine the model equations. The model l, which

contains only the difficulty index parameter (b) is known as the Rasch Model or 1 logistic parameter (1 PL), while the model containing the difficulty index (b) and discriminant index (a) is called the 2 PL model. If it contains the difficulty index (b), discriminant index (a) and pseudo-guessing (c) are called 3 PL models (Retnawati, 2016).

The desired model must have properties which the characteristics of the items do not depend on the sample group or students who are subjected to the test, the score that states the student's ability does not depend on the test, the model is stated in the item level (not at the level of a set of tests), the level model does not require parallel tests to calculate the reliability coefficient and the model provides an appropriate measure for each ability score. At the same time, the reliability of the test is expressed in the test information function, which is the sum of the information functions of all test items (Hambleton et al., 1991). If in the Classical Test Theory student's abilities are expressed in the form of the number of correct answers, which depend on the level of difficulty of the test, then in the Item Response Theory, the estimation of student's abilities is conducted by using the Maximum Likelihood Estimation (MLE) which involves test items that have been calibrated (Mardapi, 2012).

Based on these reasons, it is necessary to analyze the items using the IRT approach to estimate the reliability of the test and item parameters on the Mid-Semester Examination test and estimate student abilities. This analysis aims to provide information to schools regarding the characteristics of the questions that have been tested, estimate student abilities in the distance learning period and provide suggestions to question developers for improvement of the next questions. A further purpose of this analysis is to improve the distance learning strategy post-Middle Semester Exam and select the proper item inserted into the question bank.

2. METHODS

This research is quantitative descriptive research conducted at about Even Semester Mid-Exam in Vocational High School in Kalasan for 10th-grade mathematics courses. The data comes from the answer sheets for all students of 10th grade, which are 373 students. The data collected is in the form of student responses to the Even Semester Mid-Exam, which consists of 30 multiple choice questions with 5 options. Tests carried out with other modes of CBT (Computerized Based Test) remotely by using Moodle platform.

The analysis begins with the proof of unidimensionality assumption using factor analysis based on scree-plot and Eigenvalue on the unexplained variance (Retnawati, 2016). The factor analysis was preceded by an analysis of the adequacy of the sample based on the Chi-square value on the Bartlett test with the help of SPSS. At the same time, the assumption of local independence is evidenced by preparing a covariance matrix with the help of Excel. The last assumption is item parameter invariance and ability invariance, where item parameter invariance is proven by estimating item parameters in students who have been grouped into two groups. While the ability invariance is proven by estimating the ability using items that have been grouped into two groups. Both item parameter invariance and ability invariance are proven through scatter diagrams by checking whether the points obtained are relatively close to the line with a slope of 1 (Retnawati, 2014).

The model fit test was used to determine which of the four models was the most suitable for estimating item parameters and abilities with the item response theory approach on this test instrument. The model fit test in this research was conducted by comparing the AIC (Akaike Information Criterion) values of the four models with the help of the R program. The most suitable model is the model that has the lowest AIC (Snipes & Taylor, 2014). Having obtained the most suitable model, the next step is to estimate the item parameters and student abilities by using those models. The estimation results are then interpreted by categorizing them according to a certain range. In addition to the two estimates, the value of the item information function, the test information function and the Standard Error of Measurement (SEM) was also calculated to determine the reliability of the test.

3. FINDINGS AND DISCUSSION

Unidimensionality Assumption Test

The results of the sample adequacy analysis with the Bartlett test produce the following outputs.

Table 1 KMO and Bartlett'sTest Results

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,868
Bartlett's Test of Sphericity	Approx. Chi-Square	2066,262
	df	435
	Sig.	,000

Based on Table 1, the obtained Chi-Square is 2066.262 with 435 degrees of freedom and the significance value of 0.00, which means that the significance of less than 0.05. With a level of significance of 5%, it can be concluded that the samples of 373 students is sufficient.

Unidimensionality assumption test using EFA (Exploratory Factor Analysis) with the help of SPSS produces a scree plot of Eigenvalues .

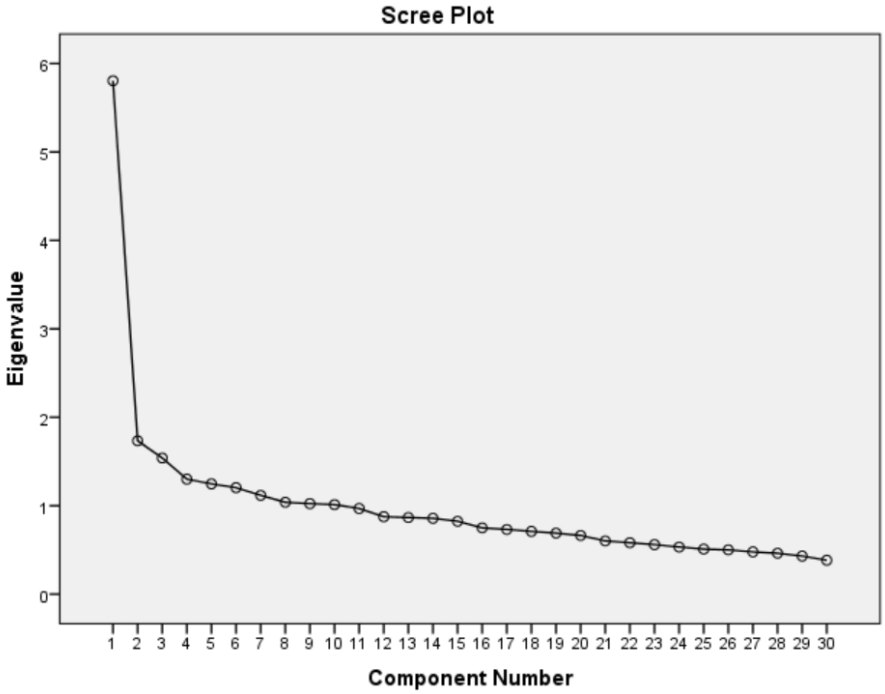

Figure 1 Scree Plot of Eigen Values of EFA Results

The scree plot is a plot between the eigenvalues and the number of main components formed and serves to determine the number of main components by taking into account the existing elbow faults (Tulak et al., 2017). Figure 1 shows that there are 10 factors that have an Eigen value of more than 1. However, there is only one dominant steepness so that it can be interpreted that in this instrument , there is one dominant factor with an Eigen value of more than 5. Based on these results, it can be concluded that this instrument is unidimensional. These results mean that each student's performance is assumed to be governed by a single factor known as ability (Eleje & Onah, 2018).

Model Fit Test

This model fit test aims to determine which analysis model best fits the data, whether the 1 Parameter Logistics (PL) model, 2 PL, 3 PL or 4 PL. If data fits into a model, then the item will behave consistently as expected by the model. There are several ways to test the model's fit, one of which is based on the AIC (Akaike Information Criterion) value. This AIC value calculates the balance between the magnitude of the likelihood and the number of variables in the model. The most suitable model is the model that has the lowest AIC (Snipes & Taylor, 2014). The calculation of the AIC value in this research was conducted with the R program, where the smallest AIC was the AIC of the 2 PL model, which was 12229.01. The model fit test in this research was conducted at the beginning after the unidimensionality test because the results of the analysis of this suitable model would then be used for the local independence and invariance assumption test.

Local Independence Assumption Test

There are two types of local independence: local independence to student responses and local independence to items. Local independence on student responses means that there is no correlation between correct or not one student in answering an item and correct or not another participant in answering the same item. This local independence on student responses can be proven through a covariance matrix that aims to see whether participants' ability in the same group is independent of the item. While local independence to items means no correlation between correct or not a student in answering an item and correct or not that student in answering other items. This means that the probability of the student answering the item correctly is not affected by the answer that the student to the other items in the test (Purnama & Alfarisa, 2020). Local independence of this test item can also be detected by proving the unidimensionality assumption (Mars, 2010 in Retnawati, 2014).

The proof of the assumption of local independence on student responses is conducted by using a covariance matrix through several steps. Students are sorted based on their ability from highest to lowest in the first step, then divided into 10 groups. The ability used in sorting is the ability based on the Maximum Likelihood Estimation (MLE) value with the most suitable analytical model, which in the case of this instrument is the 2 PL model. The resulting covariance matrix with the help of Excel is as follows figure 2.

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.47	0.09	0.04	0.05	0.04	0.04	0.06	0.08	0.09	0.15
K2		0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.04
K3			0.00	0.01	0.00	0.00	0.01	0.01	0.01	0.02
K4				0.01	0.01	0.01	0.01	0.01	0.01	0.02
K5					0.01	0.01	0.01	0.01	0.01	0.02
K6						0.01	0.01	0.01	0.01	0.02
K7							0.01	0.01	0.02	0.03
K8								0.02	0.02	0.03
K9									0.03	0.04
K10										0.08

Figure 2 Covariance Matrix

Based on the matrix above, the covariance between one group and another is shown in the elements outside the main diagonal where the value is close to 0. Thus, it can be concluded that there is no correlation between correct or not one student in answering an item and other participants in answering these items so that local independence on student responses is met. This is in accordance with the opinion of Hambleton & Swaminathan (1985) that if the covariance value of the student's ability group is close to zero, then the local independence assumption test has been fulfilled. Support for this opinion is also expressed by (Ojerinde, 2013), namely that local independence does not mean

that the items are not correlated with each other, but the performance on different items is independent but depends on the ability of students.

Parameter Invariance Assumption Test

Two invariances must be proven in fulfilling this assumption, namely the item parameter invariance and the ability parameter invariance. Item parameter invariance means that item parameters will not change if it is done by groups of students with different abilities. While the invariance of the ability parameters means that a student's ability will not change because of taking tests with different levels of difficulty (Retnawati, 2014).

Ability invariance test is conducted by estimating the ability of all respondents with response data from odd and even items separately. So this process requires two data, namely data containing the responses of all participants to odd items and data containing responses from all participants to even items. Then, estimation of the ability of each group was conducted by R program uses this data to produce two sets of the ability were then connected into a scatter plot with the help of Excel. Ability estimation was conducted using the Maximum A Posterior (MAP) method because when using the MLE estimation method, several respondents' abilities did not converge. This is because the number of items used to estimate the ability is half of the total items, which affects the estimate's stability. The effect of the number of tests on the stability of the ability estimate is in line with the results of research by Falani & Kumala (2017), which concludes that the longer the test used, the more accurate the estimation of the 2PL model's ability parameters. The scatter plot of the ability estimation with odd and even items produced in this research follows figure 3.

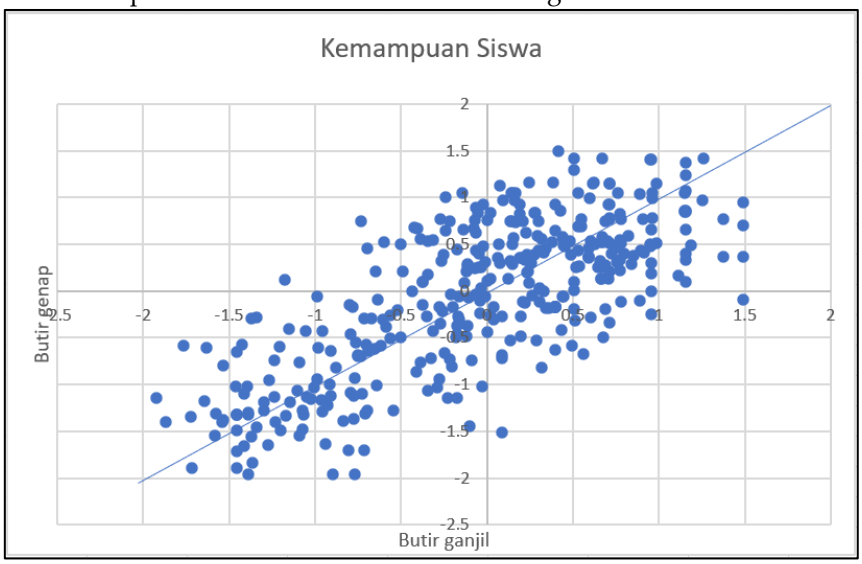

Figure 3 Scatter Plot Ability Invariance

Figure 3 above can be interpreted that the students' ability either is estimated through the odd or even items approaching the same point as the plot close to the line y = x. This shows that students' abilities do not change just because they work on items with different strengths and different levels of difficulty. Thus the invariance of the capability parameter is proven.

The next proof is the proof of the item parameters invariance assumption which includes the discriminant index parameter and the difficulty index parameter because the estimation model is 2PL. Item parameter invariance is done by dividing the respondents into two groups, namely the upper serial number group and the lower serial number group. Then, the parameter estimation of all items is conducted by using the first group of respondents, followed by the estimation of all items parameters using the second group of respondents. Two sets of discriminant index parameters and two sets of difficulty index parameters are produced. The all items discriminant index

produced by the first estimate is paired with the whole items discriminant index produced by the second estimate. With the help of Excel, the following scatter plot can be generated.

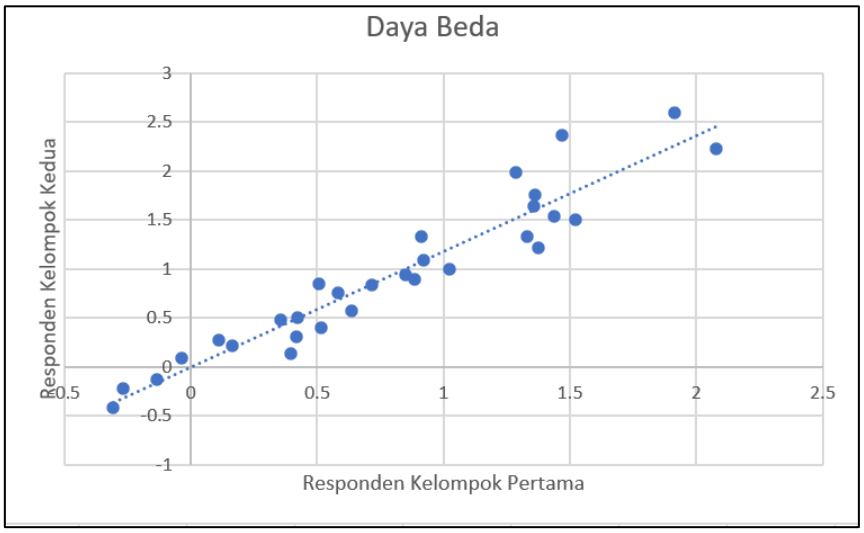

Figure 4 Discriminant Index Invariance Scatter plot

The points in the scatter plot in Figure 4 above are close to the y = x line so it can be said that the discriminant index estimated using the first group of respondents is almost the same as the discriminant index estimated using the second group of respondents. This indicates that the discriminant index does not change even though the test is solved by groups of respondents with different levels of ability. Thus the invariance of the discriminant index parameters is proven.

Proof of difficulty index invariance is done by pairing all items estimated using the first group of respondents with the difficulty index of all items estimated using the second group of respondents. The resulting scatter plot is as follows figure 5.

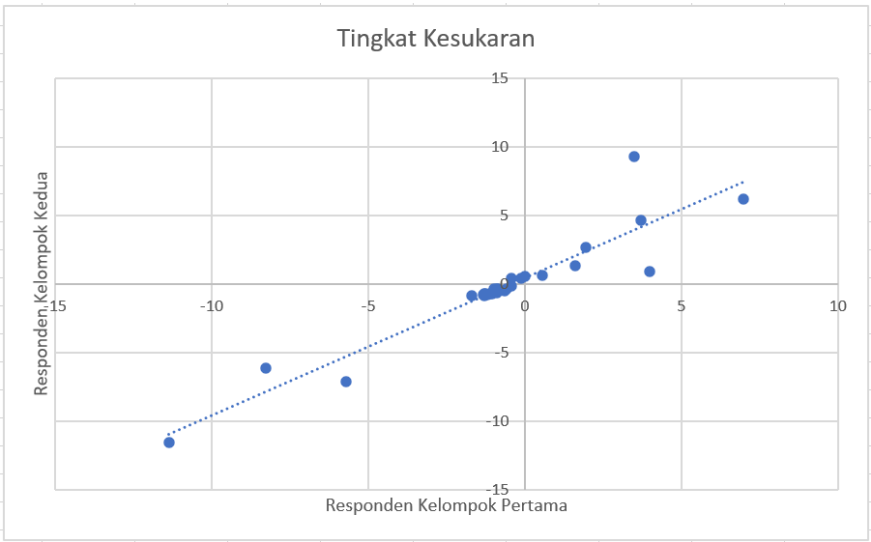

Figure 5 Difficulty Invariance Scatter Plot

Based on the scatter plot in Figure 5 above, it can be seen that the distribution of the resulting plot is close to the y = x line. This shows that the difficulty index estimated using the first group of respondents is almost the same as that of the second group of respondents. It can also be interpreted

that the difficulty index does not change even though groups of respondents do it with different abilities. Thus the invariance of the difficulty index parameter is proven.

Item Parameter Estimation

The estimation of item parameters and the Chi-Square significance value generated through the 2PL model with the help of the R Program is as follows table 2.

Table 2 Estimation Results of Item Parameters

Parameter	Item Statistics	Interpretation	Total Items	Percentage
Difficulty Index (b)	b < -2.00	Very Easy	3	10%
	-2.00 < b < -1.0 0	Easy	0	0%
	-1.00 < b < 1.00	Moderate	21	70%
	1.00 < b < 2.00	Difficult	1	3.33%
	b > 2.00	Very Difficult	5	16.67%
Discrimination Index (a)	a > 0.5	Good	20	66.67%
	0.00 < a < 0.5	Very Good	7	23.33%
	a < 0.00	Not Good	3	10%
Chi_Square (p)	p ≥ 0.05	Good because it fits the model	28	93.33%
	p < 0.05	Undescriptable because it doesn't fit the model	2	6.67%

Based on the difficulty index table above, it can be seen that the distribution of item difficulty indexes is not balanced. This can be seen from the proportion of easy items of 0%, less than the number of items with a very easy difficulty level which reaches 10%. The imbalance can also be seen from the proportion of items with a level of difficulty that is difficult and very difficult, where the number of very difficult questions is actually more than the number of difficult items. This proportion is not ideal. One of the basic references in determining the proportion of the number of questions in the difficult, medium, and easy categories is a balance based on the direction of the normal curve (Sudjana, 2012). To achieve the normal curve balance, this test instrument needs to increase the proportion of easy and difficult questions and reduce the proportion of very easy and very difficult questions. In addition, it is also necessary to consider that the proportion of easy questions is not much different from the proportion of difficult questions and the proportion of very easy and very difficult questions.

There are three items that have bad criteria because they have a negative discriminant index and the difficulty index is less than -2, namely items number 15, 25 and 29. Item number 15 is a question regarding the conversion of an angle in polar coordinates into Cartesian coordinates, item number 25 is a question regarding the formula for the area of an octagon if the symbol for the radius of the outer circle is known and item number 29 is a question about trigonometric ratios of double angles . The negative discriminant index indicates that the higher the student's ability, the lower the chance of answering correctly. The difficulty index of less than -2 indicates that the question is very easy. Based on these two things, it can be concluded that these three questions can be answered very easily by students with lower abilities than students with higher abilities. Thus, these three items cannot be included in the question bank because they are not good questions.

The possible reason for these characteristics is that the three materials have not been accepted by students who actually have higher abilities but have been accepted by students with lower abilities. 10th grade is taught by four different mathematics teachers so that it is also possible to differ in the sub-materials that are conveyed and their depth. Based on this phenomenon, the coordination

between teachers at the beginning of the semester when compiling Syllabus, Semester Programs and Learning Plans (RPP) needs to be improved to synchronize the learning process.

Two items do not fit the 2PL model, namely numbers 1 and 28. This means that the response data from these two items do not behave as expected by the 2PL model. Item number 1 is about the value of tangen an angle in a right triangle while item number 28 is about the value of sinus a half-angle α. Because these two items do not fit the model, the discriminant index and the difficulty index cannot be explained. Thus, these two questions cannot be entered into the question bank.

The question bank has an important role in the development of quality exam questions. A question bank is not just a collection of questions, but a question bank refers to a process collected, monitored and stored in a database with relevant information (classified) in order to easily search and select questions for exam purposes (Widana, 2014). According to Choppin (1976), the question bank must be understood as a collection of test items organized in the form of a catalogue like the arrangement of books in a library, only that the question bank is equipped with calibrated data and item characteristics. Thus, when the test is composed of items taken from the question bank, this calibration can be used to determine the test's psychometric properties. In relation to item response theory, Yahya Umar in Suyata et al. (2010) stated that in a question bank developed with item response theory, tests could be made more flexible and appropriate because the characteristics of test items in item response theory do not depend on the characteristics of a student at the time of calibration. In addition, students' ability can be known and compared because the ability parameters can be estimated on the same scale. Thus, item analysis with the item response theory approach in this study will support the development of a question bank with the aim of improving the quality of mathematical assessment.

Ability Estimation

The estimated ability of 373 students using the Maximum Likelihood Estimation (MLE) method through the 2 PL model with the help of Program R is in the range of -4.69 logit and 4.09 logit with an average of 0.05 logit, median of 0.12 logit and mode of 0.09 logit. Further student abilities are presented in the description table as follows.

Table 3 The Result of Student's Abilities Categories

Ability (θ)	Interpretation	Total Student	Percentage
$\theta > 3.00$	Very High	5	1.34%
$1.00 < \theta < 3.00$	High	57	15.28%
$-1.00 < \theta < 1.00$	Moderate	237	63.54%
$-3.00 < \theta < -1.00$	Low	73	19.57%
$\theta < -3.00$	Very Low	1	0.27%

The table above illustrates that the ability with the largest proportion is the ability with medium criteria and less than 20% with low and very low criteria. This composition still needs to be improved through various efforts to improve learning and increase student learning motivation during the pandemic. For this reason, it is also necessary to study the theories and results of previous research regarding mathematical abilities.

Several factors influence mathematical ability. According to Dwianjani & Candiasa (2018), the factors that contribute to mathematical problem-solving abilities include; identifying problems (identify), defining goals (define), exploring possible strategies (explore), acting on strategies (act) and looking back (look). The improvement of students' mathematical abilities can be improved through several strategies including; the use of realistic mathematics-based teaching materials (Ulandari et al., 2019), the application of project-based learning and guided discovery learning (Supriadi et al., 2018),

behaviour-based metacognition (Salam et al., 2020) and so on. Another point of view of improving students' abilities is the success of distance learning that is currently being implemented. Teachers and parents spearhead this effort. This is in line with the results of Syafari & Montessori's (2021) research, namely, the quality of online learning organized by teachers has a significant influence on students' learning motivation and has a significant influence on student learning outcomes.

Distance learning or distance education has different characteristics from face-to-face learning. Distance education is planned learning, which usually occurs in other places outside the teaching place. Therefore it requires special subject design techniques, special learning techniques, special methodologies of communication through various media, and special organizational and administrative arrangements (Moore & Kearley, 1996). Several factors influence the success and effectiveness of online learning. The effectiveness of online learning depends on technological factors, student characteristics and teacher characteristics. The teacher plays a central role in the effectiveness of online learning, it is not an important technology but the instructional application of technology from the teacher that determines the effect on learning, students who attend classes with instructors who have positive attitudes towards the distribution of learning and understanding of a technology will tend to produce a more positive learning (Pangondian et al., 2019).

In addition to the teacher's role, parents also have an equally important role, namely as teacher partners in the implementation of distance learning. This is because the family environment is the first social environment of students, which influences student behavior and motivation, especially during a pandemic where learning from home is an obligation. Regarding the role of parents in distance learning during this pandemic, a study conducted by Kusuma (2021) concluded that there is a significant relationship between parental attention and learning achievement and student learning discipline. Thus the attention and support of parents is very important in the effort to make distance learning success. In addition, collaboration and communication between teachers and parents is absolutely necessary to understand and provide solutions to problems that occur during distance learning so that students' abilities can be improved effectively.

Information Function and Standard Error of Measurement (SEM)

The information function is an indicator of the reliability of a test, where in classical test theory, the reliability of a test is expressed in the reliability coefficient. This is in accordance with the opinion of Mulvia et al. (2021) that the information function of item response theory can be used to determine the consistency or reliability of test items. This function can explain for both the item level of the test and the level of the test where at the item level of the test, the information function states the constancy or strength of the test items in explaining the respondent's ability or latent trait as measured by the test (Myszkowski, 2019). Retnawati (2014) states that the item information function explains one of the strength of the item on the test while SEM is a measurement error that cannot be separated from each estimate. The information function and SEM have a quadratic inverse relationship, where the larger the information function, the smaller the SEM and vice versa (Hambleton et al., 1991). The function of information and SEM in this test instrument is presented in the following figure 6.

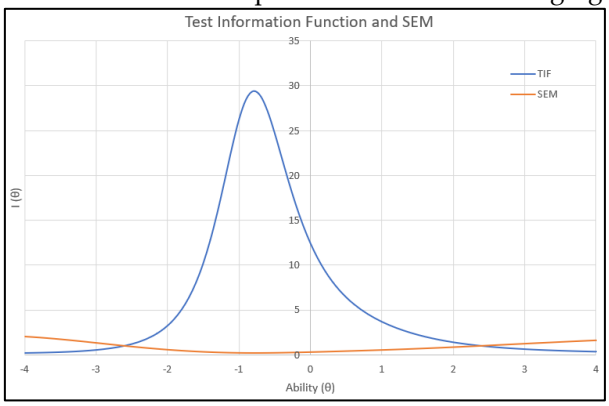

Figure 6 Test Information Function and SEM Graph

Based on the resulting graph above, the value of maximum Test Information Function (TIF) is 29,37 on the ability (θ) of -0,8 logit and Standard Error of Measurement (SEM) of 0.18 logit. This means that the test produces optimal information when used on students with an ability of -0.8 logit. In addition, according to Hambleton (in Wiberg, 2004), a good (reliable) test has a Total Information Function (TIF) ≥ 10. Thus, this instrument is a good (reliable) instrument for measuring students' abilities. The TIF and SEM curves intersect at = -2.6 and 2.4, which means that this test is overall suitable for students with abilities between -2.6 and 2.4 logit because in this range, the value of the information function exceeds the error of measurement. This range also shows that the instrument is able to measure student's abilities with a fairly wide range.

CONCLUSION

Based on the findings and discussion above, it can be concluded that this test instrument consists of 25 items that are good and deserve to be included in the question bank for further use and 5 items that need to be replaced or revised. This is based on the difficulty index, discriminant index and compatibility with the 2PL model. The average and mode of student's abilities are in the medium category, while the median ability of students is in the high category so it still needs to be improved through various improvements to learning strategies and environmental support. As for the function of information and SEM on this test instrument, it can be concluded that this test can be used for measurement and is suitable for test-takers with abilities between -2.21 and 1.06 logit.

Suggestions that can be put forward regarding this test instrument are 1) improvement in the distribution of difficulty index so that the proportion of difficulty index close to the normal curve, 2) relates to learning practices in schools, it is needed initial coordination between all of teacher and synergy in the learning process so that students get material with relatively the same scope and depth, 3) related to improving student abilities, it is necessary to maximize the role and strategies of teachers in learning as well as the role and support of parents for student during distance learning to improve students' abilities, 4) improving the quality of communication and synergy between parents and teachers in solving problems that arise during distance learning is absolutely necessary so that learning runs optimally, and 5) based on the weaknesses found in this instrument, the suggestion for researchers and developers of the next question is in the planning of questions, it is better to involve items whose characteristics are known, which are called calibrated questions or questions that have been tested before. As for the items in this instrument that have been proven to have a good level of difficulty and discriminating power, they can be used by question developers or further researchers to improve the quality of mathematics assessment instruments. And vice versa, the questions that are not good can be revised or replaced with new items by considering the relevant indicators of competency achievement.

REFERENCES

Azwar, S. (2016). Tes Prestasi (2nd ed.). Pustaka Pelajar.

Azwar, S. (2019). Reliabilitas dan Validitas (IV). Pustaka Pelajar.

Choppin, B. (1976). Development in Items Banking. National Standards of Attainment in Schools.

Direktorat Pembinaan Sekolah Menengah Kejuruan Kemdikbud RI. (2018). Panduan Penilaian Hasil Belajar dan Pengembangan Karakter pada Sekolah Menengah Kejuruan. Kemdikbud RI. psmk.kemdikbud.go.id

Dwianjani, N. K. V., & Candiasa, I. M. (2018). Identifikasi Faktor-Faktor yang Mempengaruhi Kemampuan Pemecahan Masalah Matematika. Jurnal Matematika Dan Pendidikan Matematika, 2(2). https://doi.org/https://doi.org/10.25217/numerical.v2i2.276

Eleje, L. ., & Onah, F. E. (2018). Comparative Study of Classical Test Theory and Item Response Theory Using Diagnostic Quantitative Economics Skill Test Item Analysis Results. European Journal of Educational & Social Sciences, 3(1), 71–89.

Falani, I., & Kumala, S. A. (2017). Kestabilan Estimasi Parameter Kemampuan Pada Model Logistik Item Response Theory Ditinjau dari Panjang Tes. Jurnal Susunan Artikel Pendidikan (SAP), 2(2). https://doi.org/10.30998/sap.v2i2.2028

Hambleton, R. K., & Swaminathan, H. (1985). Items Response Theory: Principles and Application. Kluwer-Nijjhoff Publish.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). Fundamentals of Item Response Theory. Sage Publications.

Istiyono, E. (2018). Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika Dengan Teori Tes Klasik dan Modern (First). UNY Press.

Kusuma, Y. Y. (2021). Analisis Hubungan Perhatian orang Tua Dengan Prestasi Belajar Pada Masa Pandemi Covid - 19. Jurnal Pendidikan Dan Konseling (JPDK), 3(1). https://doi.org/https://doi.org/10.31004/jpdk.v3i1.1384

Mardapi, D. (2012). Pengukuran Penilaian dan Evaluasi Pendidikan. Nuha Medika.

Moore, M., & Kearley, G. (1996). Distance education: A systems view. Wadsworth.

Mulvia, R., Ramalis, T. R., & Efendi, R. (2021). Mendeteksi Keajegan Butir Tes dengan Fungsi Informasi. Jurnal Pendidikan Indonesia, 2(1). https://doi.org/https://doi.org/10.36418/japendi.v2i1.66

Myszkowski, N. (2019). Development of the R library "jrt" : Automated Item-Response Theory procedures for judgment data and their application with the Consensual Assessment Technique. Psychology of Aesthetics Creativity and The Arts. https://doi.org/http://dx.doi.org/10.1037/aca0000287

Ojerinde, D. (2013). Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of The Comparability of Item Analysis Result. Semantic Scholar. https://www.semanticscholar.org/paper/Classical-Test-Theory-(-CTT-)-VS-Item-Response-(-)-Ojerinde/aa27637d699b7e26ee1f7578d0fe2b8173ad769f

Pangondian, R. A., Santosa, P. I., & Nugroho, E. (2019). Faktor - Faktor Yang Mempengaruhi Kesuksesan Pembelajaran Daring. Seminar Nasional Teknologi Komputer & Sains (SAINTEKS), 56–60.

Purnama, D. N., & Alfarisa, F. (2020). Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi SMK Berdasarkan Teori Tes Klasik dan Teori Respons Butir. Jurnal Pendidikan Akuntansi Indonesia, 18(1), 36–46. https://doi.org/http://dx.doi.org/10.21831/jpai.v18i1.31457

Retnawati, H. (2014). Teori Respons Butir dan Penerapannya. Parama Publishing.

Retnawati, H. (2016). Validitas Reliabilitas & Karakteristik Butir. Parama Publishing.

Salam, M., Misu, L., Rahim, U., Hindaryatiningsih, N., & Ghani, A. R. A. (2020). Strategies of Metacognition Based on Behavioural Learning to Improve Metacognition Awareness and Mathematics Ability of Students. International Journal of Instruction, 13(2), 61–72. https://doi.org/https://doi.org/10.29333/iji.2020.1325a

Snipes, M., & Taylor, D. C. (2014). Model selection and Akaike Information Criteria: An Example from Wine Ratings and Prices. Wine Economic and Policy, 1. https://doi.org/http://dx.doi.org/10.1016/j.wep.2014.03.001

Sudjana, N. (2012). Penilaian hasil Proses Belajar Mengajar. Remaja Rosdakarya.

Supriadi, N., Syazali, M., Lestari, B. D., Dewi, E. S., Utami, L. F., Mardani, L. A., & Putra, F. G. (2018). The Utilization of Project Based Learning and Guided Discovery Learning: Effective Methods to Improve Students' Mathematics Ability. Al-Ta'lim Journal, 25(3). https://doi.org/https://doi.org/10.15548/jt.v25i3.487

Suyata, P., Mardapi, D., & Kartowagiran, B. (2010). Identifikasi Need Assessment:Studi Awal Model Pengembangan Bank Soal Berbasis Guru Di Provinsi DIY. Jurnal Kependidikan, 40(1). https://doi.org/https://doi.org/10.21831/jk.v40i1.512

Syafari, Y., & Montessori, M. (2021). Analisis Pembelajaran Daring Terhadap Motivasi Belajar Dan Prestasi Belajar Siswa Dimasa Pandemi Covid-19. Jurnal Basicedu, 5(3). https://doi.org/https://doi.org/10.31004/basicedu.v5i3.872

Tulak, E. T., Raupong, & Ilyas, N. (2017). Penerapan Regresi Robust Principal Component Analysis Pada Data yang Mengandung Multikolinieritas dan Outlier [Universitas Hasanudin]. http://digilib.unhas.ac.id/opac/detail-opac?id=38231

Ulandari, L., Amry, Z., & Saragih, S. (2019). Development of Learning Materials Based on Realistic

Mathematics Education Approach to Improve Students' Mathematical Problem Solving Ability and Self-Efficacy. International Electronic Journal of Mathematics Education, 14(2), 375–383. https://doi.org/https://doi.org/10.29333/iejme/5721

Wiberg, M. (2004). Classical Test Theory vs. Item Response Theory. Umea Universitet.

Widana, I. W. (2014). Pengembangan Bank Soal. Jurnal EMASAINS, 3(2), 186–197.

Worten, B. R., & Sanders, J. R. (1987). Educational Evaluation: Theory and Practice. Wadsworth.

This page is intentionally left blank