Educational Administration: Theory and Practice
2024,30(4), 3799-3810
ISSN:2148-2403
https://kuey.net/
Research Article

Analysis of Item Characteristics in Elementary School Mathematics Reasoning Assessment Using Item Response Theory with the Generalized Partial Credit Model

Noening Andrijati[1*], Heri Retnawati[2], Sudiyatno[3]

[1*]Doctoral Student, Educational Research and Evaluation Study Program, Postgraduate Program, Yogyakarta State University, Yogyakarta, Indonesia
[2]Lecturer of Mathematics Education Study Program, Faculty of Mathematics and Natural Sciences, Yogyakarta State University, Yogyakarta, Indonesia
[3]Lecturer of Mechanical Engineering Education Study Program, Faculty of Engineering, Yogyakarta State University, Yogyakarta, Indonesia

***Corresponding Author:** Noening Andrijati
*Doctoral Student, Educational Research and Evaluation Study Program, Postgraduate Program, Yogyakarta State University, Yogyakarta, Indonesia

Citation: Noening Andrijati et Al. (2024), Analysis of Item Characteristics in Elementary School Mathematics Reasoning Assessment Using Item Response Theory with the Generalized Partial Credit Model, *Educational Administration: Theory And Practice*, *30*(4), 3799-3810
Doi: 10.53555/kuey.v30i4.2129

ARTICLE INFO

ABSTRACT

This survey-based research focuses on the analysis of item characteristics in the assessment of mathematical reasoning among elementary school students, employing the Item Response Theory (IRT) and the Generalized Partial Credit Model (GPCM). The findings are as follows: 1) Overall, the knowledge assessment items are categorized as satisfactory, with discrimination indices ranging from 0.422 to 2.401 and difficulty indices from -1.31 to 0.924 on the logit scale. A single item requires revision due to acceptable difficulty levels paired with poor discrimination. The knowledge assessment instrument provides precise information across a skill range of -2.5 to 1.7, peaking at approximately -0.8; 2) Generally, the skill assessment items are also rated as satisfactory, with discrimination indices from 0.561 to 1.554 and difficulty indices from -2.335 to 1.808. Three items necessitate corrections due to unfavorable difficulty levels despite adequate discrimination. The skill assessment instrument accurately measures abilities within a range of -3.2 to 2.6, with the most significant information at approximately 1.3; 3) Lastly, the attitude assessment items are generally good, with discrimination indices from -0.071 to 1.635 and difficulty indices from -0.588 to 1.273. Three items require adjustments due to poor difficulty levels, despite satisfactory discrimination. The attitude assessment instrument effectively gauges the range of -1.75 to 2.5 in skills, with the highest information value around 0.5. This study underscores the effectiveness of the GPCM in discerning item characteristics that are crucial for enhancing the assessment of mathematical reasoning in elementary education.

Keywords: Item response theory, Generalized partial credit model, Mathematics, Reasoning, Knowledge Assessment, Skill Assessment, Attitude Assessment.

INTRODUCTION

Reasoning is an important aspect in learning mathematics, even it becomes one of the aims of learning mathematics. As stated in Government Regulation Number 32 of 2013 about National Education Standard, that the scope of material study of mathematics includes arithmetic, geometry, and algebra aimed to develop students' logical thinking and thinking skill. The math learning goal in National Education Standard corresponding with the general goal of learning mathematics that has been formulated by the National Council of Teachers of Mathematics or NCTM (2000: 7), includes 1) problem solving skill. 2) reasoning skill, 3) communication skill, 4) connection skill and 5) representation skill.
Several research about reasoning skill of mathematics have been done, both in Indonesia and international scope. The study of Utari Sumarmo found that the score of students' reasoning skill is still low, in line with

the study of Wahyudin (2008) that reveals that one of tendencies causing the number of students is fail in mastering well the subject in mathematics is the students lack in using the logical reasoning in solving the mathematics problem given. Similarly, the study of Rif'at in Priatna (2003) give the conclusion that the lack of students' mathematics skill can be observed from their performance of reasoning, doing error in solving mathematics question is due to their error in using reasoning. The founding of these research result shows the low of students' reasoning skill effects on the weak mastering in mathematics content.

One of the studies of international level about the reasoning skill of mathematics has been conducted by the institution study The Trends in International Math and Science (TIMSS) that hold the survey to monitor the students' mathematics and science achievement of fourth grade of elementary school (age 9 to 10 years old) and second grade of junior high school (age 13 to 14 years old) all over the world. TIMSS is a continued study and conducted in every four years. This study is a long series of the study conducted by International Association for the Evaluation of Educational Achievement (IEA) to assess the achievement in education. Indonesia has participated as a participant in TIMSS in five times, such as in 1999, 2003, 2007, 2011 and 2015. On the TIMSS maintenance in 1999. 2003. 2007 and 2011 Indonesia includes the second-grade students of junior high school. Meanwhile, in 2015, Indonesia includes only the fourth-grade students of elementary school (Mullis & Martin, 2014).

The using of the right assessment model will really determine the success in access the information relating to the learning process. The selecting of assessment model should be based on the information target that have to be achieved. The information mean is the result of learning outcome achievement that achieved by students. Stiggins (1994: 67) propose five learning targets that is proper to become the basic in determining the kinds or model of assessment that will be used by the teacher. The five-learning target is corresponding with the domain measured in authentic assessment and competency of 21st Century, such as 1) knowledge, is content of subject that should be mastered by students that covers the knowledge and understanding; 2) reasoning, is a skill in using the knowledge and understanding to find and solve the problem; 3) performance skill, developing of attitude or the process of skill; 4) product, is a skill to make a real product that meets the certain standard; 5) dispositions, is the development of attitude, motivational interests and intentions supporting the successfulness of students learning at school. Therefore, reasoning includes, or mathematics reasoning is one of the learning targets from authentic assessment. The assessment of mathematics reasoning can use various assessment technique used in authentic assessment.

Several research about authentic assessment in learning mathematics have been conducted. Sujaya, Suarni, dan Candiasa (2013) have conducted the researching producing finding that the learning model of authentic assessment and achievement motivation has significant effect to the result of learning mathematics on fifth grade students of elementary school. The finding gives the implication that learning model of authentic assessment needs to be considered in the process of learning mathematics and the implementation of performance assessment should consider the students' high motivation. This result of research gives the strengthen that the authentic assessment model is needed and relevant to be applied in learning mathematics at elementary school.

The study of Badrun Kartowagiran & Amat Jaedun (2016: 139) and Rivo Eka Yuda (2016) show the similar result of study. Badrun Kartowagiran and Amat Jaedun (2016: 1 & 39) give a conclusion on the research that the teacher (implementing authentic assessment) still needs the improvement and the quality of authentic assessment implementation in junior high school and needs improvement. Outline, the conclusion is based on the condition that: 1) the assessment design stated in lesson plan is still deficient, 2) there only few teacher conducting the assessment of attitude, assessment that is integrated with learning and continued; 3) all teachers do the assessment technique of knowledge and almost all teachers do the assessment technique of skill, but it is not variative. Meanwhile, the research conclusion from Rivo Eka Yuda (2016: 10-11) state that assessment of work method made by elementary school teacher is still in the form of a written test and assignment sheet. These both research results indicate that it needs the improvement of the quality of authentic assessment implementation in school, mainly about the variation technique or instrument used to measure the three domains of assessment (knowledge, skill and attitude).

Sri Wardhani (2010) confirm that there still many teachers that have been not skilled yet in developing the assessment instrument of learning result. Besides that, teachers tend to develop the assessment instrument with objective question form or essay question that usually used in test activity by written test technique. Teachers are not accustomed to developing the assessment instrument by using complex written test technique or not written test technique for instance performance test or project assignment. Moreover, teachers also have not been optimal in developing the assessment instrument oriented in mathematics subject goal, as well as the learning standard and mathematics assessment standard. The conclusion based on the result of observation in training activity or Subject Teacher Conference facility.

An essay test or assignment to measure students' performance in the dimension of knowledge and skill in mathematics learning assessment is usually arranged by using response model more than two categories (polytomous). Polytomous scoring means that assessment instrument in the form of essay test or assignment can give more information about item question characteristics and students' ability, therefore it could draw unilinear relationship between the participant of the test and skill θ and the probability of the test participant answer the item response in certain category. Therefore, the researchers used Item Response Theory (IRT).

Item Response Theory basically comes to improve the weakness in classical test theory which is group independent and item independent. In classical test theory, discrimination index, difficulty level and coefficient of reliability of the test depend on someone doing the test, besides it is affected by the question or item (Samsul Hadi, 2013: 10). Meanwhile, IRT build a model that connect the characteristics of item with participant's characteristics. For clearer, Item Response Theory is a theory how variable people and item determine the response data when someone answers the item (Umar, 1999). Item Response Theory has an excess compared with classical test theory, such as 1) it does not depend on group independent, 2) the score describes individual' skill of the test participant, 3) it emphasizes on the level of item not the test, 4) it requires the proper size for every score of ability, and 5) it does not need parallel test in determining the reliability (Hambleton & Jones 1993).

The development of probability function IRT model in dichotomous score used logistic function that consist of a) model 1-P with the parameter of difficulty level, b) model 2-P with the parameter of difficulty level and discrimination level, c) model 3-P with the parameter of difficulty level, discrimination level and guessing. Meanwhile, the Item Response Theory polytomous that is known such as Nominal Response Model (NRM), Rating Scale Model (RSM), Partial Credit Model (PCM), Graded Response Model (GRM) and Generalized Partial Credit Model (GPCM) (Demars, 2010: 22; Retnawati, 2014: 32).

GPCM is an elaboration from Partial Credit Model (Muraki, 1999). In GPCM, the difficulty level of each step is calculated to estimate the participant's ability. Scoring is not conducted directly, but by using certain methos after the estimation of item parameter is conducted. GPCM is a general form of PCM, stated in the form of mathematical that called as the function of Item Response category (Muraki & Bock, 1997:154; Retnawati, 2014:).

Based on the background of the study, this research focuses on the analysis of item characteristics of the assessment instrument of knowledge, skill and attitude by Item Response Theory Approach and by GPCM model.

RESEARCH METHOD

This research is survey research, i.e., collecting the data from a group, in this case is students, and the data is analyzed to find out the quality of item of assessment instrument of knowledge, skill and attitude.

This research is started from arrangement of test specification, indicator and item based on the construct theory that has been done. Then the instrument that has been arranged, made content validation of instrument, where the result of validation in each aspect of assessment is presented on the table below.

Table 1. The Calculation Result of Index V Aiken of Assessment Instrument of Knowledge

Kinds of instrument	Number of item	The Average of Index Aiken v	Explanation
Knowledge Assessment	9	0,876	Valid
Skill Assessment	12	0,875	Valid
Attitude assessment	15	0,881	Valid

The result of content validation is obtained based on the experts' assessment to the assessment instrument of mathematics reasoning skill developed (draft 1). Each expert gives assessment about content representation using validation sheet that has been prepared in each item test of assessment instrument. Index of validity is calculated to notice the consistency between validators by using the formulation of index Aiken V (Allen, 1985; Kumaidi, 2014; Heri Retnawati, 2015:18). The assessment used sale 4 (1 to 4) and calculated with the formulation (1) in section II before.

Table 1 is a table of calculation result of index V Aiken average in each assessment instrument. Table 1 indicates the attitude assessment that consist of nine items that have index Aiken V average which is 0.876. In the assessment instrument of skill that consist of 12 items indicate the result of index Aiken V average is 0.875. meanwhile in the assessment instrument of attitude that consist of 12 items indicate the index Aiken V average is 0.881. of three result show the valid description in each instrument.

The reliability of assessment instrument based on internal estimation consistency with coefficient Alpha from Cronbach (Cronbach's Alpha). The calculation of coefficient reliability Alpha is conducted by SPSS program version 22. The good criteria coefficient reliability is ≥ 0.7 (Nunnaly, 1981: 245. Meanwhile, Ebel & Frisbie (1991: 86) state that if the test is used as standard test, the coefficient reliability should be between 0.85 – 0.95, while if it used for classroom assessment ≥ 0.65. the result of estimation of reliability can be viewed on the table 2 bellow.

Table 2. The Result Of Reliability Estimation of Assessment Instrument

Dimension of Assessment Instrument	Reliability Coefficient	
	Construct Reliability	*Cronbach's Alpha*
Knowledge	0,93	0,986
Skill	0,96	0,978
Attitude	0,98	0,747

Based on the table 2, it is found that coefficient of Cronbach's Alpha for assessment instrument of skill reasoning of knowledge dimension is 0.968 > 0.7, so it can be stated that the instrument is reliable and considered having good consistency. This case shows that coefficient of Cronbach's Alpha for assessment instrument of reasoning skill dimension is 0.978 > 0.7, so it can be concluded that the assessment instrument of skill is reliable and considered having good consistency. Meanwhile, the coefficient of Cronbach's Alpha for assessment instrument of reasoning skill in attitude dimension is 0.747 > 0.7, therefore it is concluded that assessment instrument of attitude is reliable and considered having good consistency. The decision of this analysis result is based on the category of Cronbach's Alpha reliability and good criteria of reliability coefficient which is minimum 0.7 (Nunnaly, 1981: 254).

From the table 2, it can be found out the quantity of reliability coefficient from assessment instrument of knowledge, skill and attitude. The coefficient of construct reliability of assessment instrument of knowledge, skill and attitude is in a row 0.93, 0.96, and 0.98, as well as of the three is > 0.7, therefore it can be concluded that the assessment instrument of reasoning skill of knowledge, skill and attitude dimension is reliable and having good consistency.

After found out how the validity of instrument content and the result of estimation of reliability, the next step is evaluating the instrument to find out the item characteristics. The characteristics of item question of assessment instrument of mathematics reasoning skill of knowledge and skill dimension is determined by using IRT approach of Polytomous GPCM model help program R MIRT package. The information obtained through this analysis includes item parameter estimation, ability (θ), curve of item characteristics, curve of item information and assessment instrument as well as standard error of measurement.

RESULT OF RESEARCH

Assessment Instrument of Reasoning Skill of Knowledge Dimension

The result data of mathematics reasoning skill measurement that will be analyzed by IRT model GPCM, first conducting the item fit examination. Item fit examination is used to find out whether 9 item question developed is fit with the GPCM model or 2PL used. The summery of analysis result of item fit for assessment instrument is presented on table 3 below.

Table 3. The Result of Item Fit Test of Knowledge Assessment Instrument

Item	**χ^2 (*Chi-Square*)**				**Explanation**
	Statistics	**Df**	**RMSEA**	***p- value***	
1A	33,363	33	0,009	0,450	Fit
2A	19,724	23	0,000	0,658	Fit
3A	12,385	13	0,000	0,496	Fit
1B	26,077	19	0,054	0,128	Fit
2B	25,676	38	0,000	0,936	Fit
3B	33,915	23	0,061	0,066	Fit
1C	38,664	28	0,055	0,086	Fit
2C	14,637	15	0,000	0,478	Fit
3C	13,759	17	0,000	0,684	Fit

The match of item question can be viewed from the value of p (p-value). If p-value the result of analysis is > 0.05, so the item fit with the model used. Based on the analysis result of item fit on the table 40 and the criteria of item fit, it can be found that all items have p- value > 0.05. Therefore, it can be concluded that all item of assessment instrument of reasoning skill of mathematics in knowledge dimension is fit with the model 2PL used. The measurement model of assessment instrument of reasoning skill in knowledge dimension is presented on the figure 1 below.

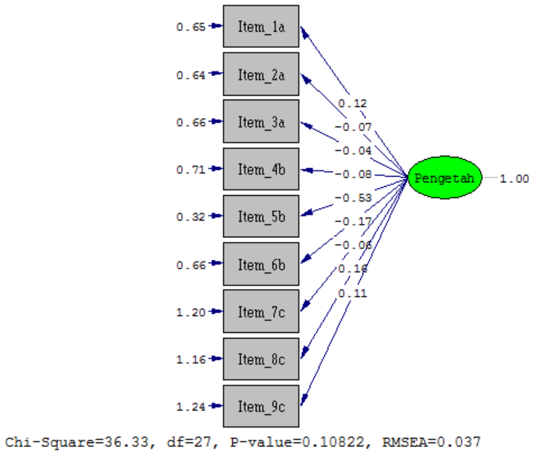

Figure 1. Measurement Model of Knowledge Assessment

After all item is stated fit with the model, next is conducted the item parameter estimation of knowledge assessment. Item parameter estimation is conducted to find out the quality of item question. Item parameter estimated includes the difficulty level (bi) and discrimination (ai). the summary of item parameter estimation result of knowledge assessment is presented on table 4 below.

Table 4. The Result of Item Parameter Estimation of Knowledge Assessment

No. Item	**Discrimination(a_i)**		**Difficulty Level(b_i)**						**Conclusion**
	(a_i)	Explanation	(b_1)	(b_2)	(b_3)	(b_4)	Average (b_i)	Explanation	
1A	0,662	Good	-0,032	-0,814	-0,041	1,75	0,21575	Good	Good
2A	1,616	Good	-0,71	-0,076	0,447	1,342	0,25075	Good	Good
3A	1,596	Good	-1,562	-2,011	-0,917	-0,748	-1,3095	Good	Good
1B	1,587	Good	-0,955	-0,152	-1,036	0,114	-0,50725	Good	Good
2B	0,422	Good	-1,742	-1,744	0,335	0,381	-0,6925	Good	Good
3B	1,451	Good	-1,586	-1,177	-0,236	0,221	-0,6945	Good	Good
1C	0,79	Good	-1,133	-0,447	0,511	1,286	0,05425	Good	Good
2C	2,401	Bad	0,432	0,76	1,059	1,445	0,924	Good	Corrected
3C	1,542	Good	-2,37	-1	-0,652	-0,822	-1,211	Good	Good

The data on the table 4 shows that 8 (item number 1A, 2A, 3A, 1B, 2B, 3B, 3A, and 3C) have discrimination index in range 0 – 2, therefore it can be stated that of 9 items have good discrimination. Meanwhile, item 2C has discrimination index, which is 2.401, the discrimination is out of range 0-2, therefore it can be stated that the item is bad. The difficulty level in table 41 is showed by 4 categories such as b1, b2, b3, and b4. The difficulty level in each category gives different result, but if it is viewed based on the average, it seems that all items have difficulty level in range -2 - +2. Therefore, it can be concluded that all items question has good item test category.

Referring to the analysis result of discrimination and difficulty level estimation of item can be understood that all items have good difficulty level and discrimination except item 2C. item 2C has good difficulty level but bad discrimination. Based on the result, it can be conducted the improvement in item 2C (Number 2 in Basic Competency 3.7.)

The strength of an item on the assessment instrument set, selection of test item, and comparison of assessment instrument set in IRT is explained by the item information function. The item information function states the strength or item contribution of knowledge assessment instruments in revealing the developed reasoning skill. Meanwhile, the test information function is the number of the information function for each item. The value of the information function of the assessment instrument will be high if the items of instrument arrangement have a high information function. The graph of the information function and the measurement standard error of the knowledge assessment instrument is depicted in Figure 2 below.

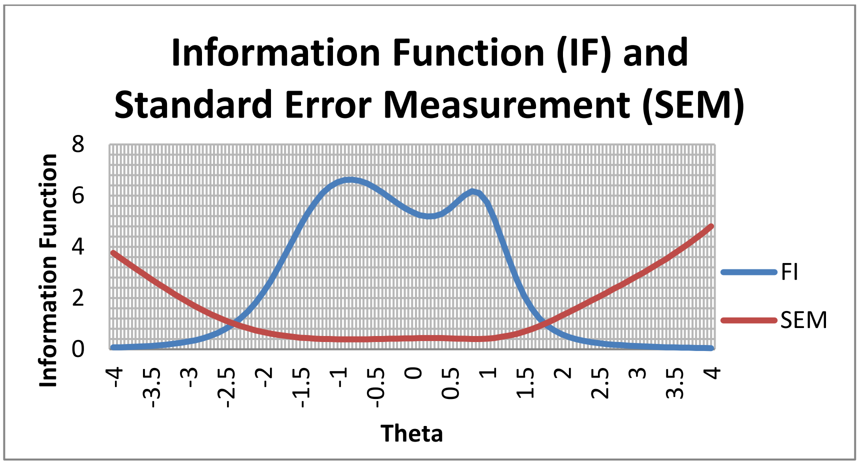

Figure 2. Graph of Information Function and Error Measurement of Knowledge Assessment Instrument

Figure 2 shows that the assessment instrument provides maximum value information which is 6.624 in θ is about -0.8 and has error measurement which is 0.388. The value of the information function obtained is in the range of -2.5 to +1.7. This case indicates that the knowledge assessment instrument is suitable for students with abilities from -2.5 to +1.7 (low to high ability).

Assessment Instrument of Mathematics Reasoning Skill in Skill Dimension

The data which is in the form of responses to the skill assessment instrument that will be analyzed using IRT of GPCM model, first checking the Item Fit. Item Fit examination is conducted to find out whether the 12 items question developed is fit with the GPCM model (2PL used). The summary of analysis result of the item fit for skill assessment instrument is presented in Table 5 below.

Table 5. The Test Result of Item Fit of Skill Assessment Instrument

Item	**χ^2 (*Chi-Square*)**				**Explanation**
	Statistic	**Df**	**RMSEA**	***p- value***	
1A	8,218	4	0,091	0,084	Fit
2A	10,724	6	0,079	0,097	Fit
3A	3,851	3	0,047	0,278	Fit
4A	5,395	5	0,025	0,370	Fit
5B	1,916	5	0,000	0,861	Fit
6B	9,563	6	0,068	0,144	Fit
7B	8,279	4	0,092	0,082	Fit
8B	1,358	5	0,000	0,929	Fit
9C	11,318	7	0,07	0,125	Fit
10C	5,610	3	0,083	0,132	Fit
11C	2,220	3	0,000	0,528	Fit
12C	9,387	5	0,083	0,095	Fit

The suitability of an item can be viewed from the p-value. If the p-value of the analysis is > 0.05, then the item fits the model used. Based on the result of the item fit analysis in Table 43 and the item fit criteria, it can be found out that all items have p-value > 0.05. Therefore, it can be concluded that the overall items of the mathematical reasoning skill assessment instrument on the skill dimensions is fit with the 2PL model used. The measurement model of the reasoning skill of assessment instrument on the skill dimension is presented in Figure 3 below.

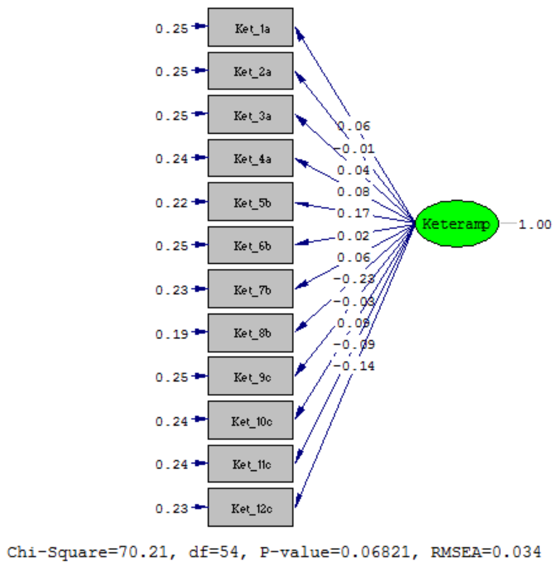

Figure 3. Measurement Model of Skill Assessment

After all the items are stated fit with the model, the next step is to estimate the item parameters of skill assessment. The estimated item parameter includes the level of difficulty (bi) and discrimination (ai). The summary of the estimation results of item parameters of the mathematical reasoning skill of assessment instrument on skill dimension is presented on Table 6 below.

Table 6. The Estimation Result of Item Parameter of Skill Assessment Instrument

Item	Discrimination		Difficulty Level		Conclusion
	a	Explanation	b	Explanation	
1A	0,678	Good	-2,128	Bad/ Low	Corrected
2A	1,025	Good	-0,146	Good	Good
3A	0,561	Good	-2,335	Bad/Low	Corrected
4A	1,145	Good	1,335	Good	Good
5B	1,554	Good	1,234	Good	Good
6B	0,858	Good	-1,526	Good	Good
7B	0,794	Good	-0,359	Good	Good
8B	0,816	Good	0,589	Good	Good
9C	0,694	Good	1,808	Good	Good
10C	1,352	Good	-1,202	Good	Good
11C	0,708	Good	-2,277	Bad/Low	Corrected
12C	0,591	Good	-1,851	Good	Good

The data in Table 6 shows that all items (12 items) have discrimination index in the range of 0 - 2, so it can be stated that all items have good discrimination. According to estimation result of difficulty level item, obtained information that 9 items question of skill assessment have difficulty index item which is in good category since they are in the range of -2 to 2, i.e., number 2A, 4A, 5B, 6B, 7B, 8B, 9C, 10C, and 12C. Meanwhile, items numbered 1A, 3A, and 11C have difficulty index item of -2.128, -2.335, and -2.277. The three items have a low item of difficulty level category.

Referring to the analysis result of the estimation of discrimination and difficulty level of item can be conceived that all items have good discrimination and difficulty level except for item 1A, 3A, and 11C. although the item 1A, 3A, and 11C have bad difficulty level, but they have good discrimination, therefore it will be conducted the improvement on item 1A, 3A, and 11C.

Furthermore, the researcher will reveal the information function of the skill assessment instrument. The representation of information function of assessment instrument can be observed through the graph of information function and error measurement on figure 4 below.

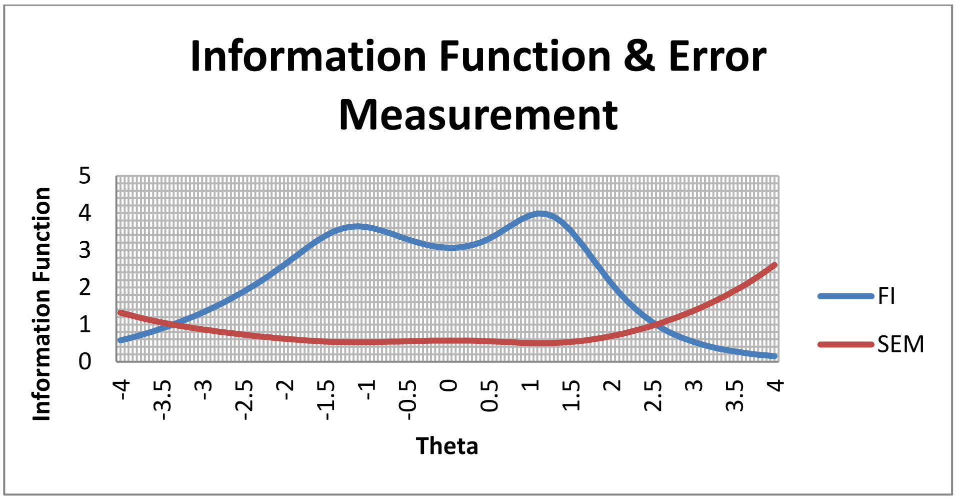

Figure 4. Information Function & Error Measurement

Figure 4 indicates that the skill assessment instrument acquires maximum information which is 3.988 on theta which is about 1.3 and have error measurement 0.5000. the value of information function obtained is on the range of -3.2 to +2.6. this case indicates that the skill assessment instrument is suitable for students with the ability -3.2 to +2.6 or students with low to high skill.

The Assessment Instrument of Reasoning Skill of Mathematics of Attitude Dimension

The data which is the form of response to attitude assessment instrument will be analyzed by IRT model GPCM, first checking the item fit. Checking the item fit is conducted to find out whether 15 items of statement developed is fit with the model GPCM (2PL used). The summary of analysis result of item fit for attitude assessment instrument is presented on the table 7 below.

Table 7. the Test Result of Item Fit of Attitude Assessment Instrument

Item	χ^2 (*Chi-Square*)				Explanation
	Statistics	df	RMSEA	*p-value*	
Sik_01	26,663	30	0	0,641	Fit
Sik_02	21,315	19	0,031	0,32	Fit
Sik_03	11,083	18	0	0,891	Fit
Sik_04	18,352	16	0,034	0,304	Fit
Sik_05	24,827	17	0,06	0,099	Fit
Sik_06	13,519	14	0	0,486	Fit
Sik_07	12,959	20	0	0,879	Fit
Sik_08	23,566	32	0	0,86	Fit
Sik_09	10,93	9	0,041	0,281	Fit
Sik_10	15,785	15	0,02	0,396	Fit
Sik_11	28,763	26	0,029	0,322	Fit
Sik_12	11,005	14	0	0,686	Fit
Sik_13	22,553	21	0,024	0,368	Fit
Sik_14	26,989	30	0	0,624	Fit
Sik_15	23,404	24	0	0,496	Fit

The match of item question can be viewed from p value. If p value of analysis result is > 0.5, the item is fit with the model used. Based on the analysis result of item fit on table 44 and the item fit criteria, it can be found out that all items have p value > 0.5.

Therefore, it can be concluded that all items of assessment instrument of reasoning skill of mathematics in attitude dimension is fit with the model 2PL used. The model of assessment instrument of reasoning skill in attitude dimension is presented on the figure 5 below.

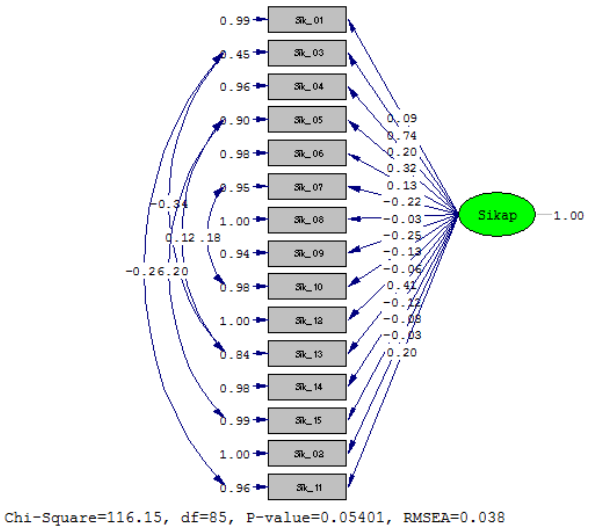

Figure 5. Attitude Assessment Instrument Model

After all items is stated fit with the model, the next step is estimating the parameter of question item of attitude assessment. The item parameter estimated includes the difficulty level (bi) and discrimination (ai). The summary of estimation result of item parameter of assessment instrument of reasoning skill mathematics in attitude dimension is presented on the table 8 below.

Table 8. The estimation result of Item Parameter of Attitude Assessment Instrument

Item	Discrimination		Difficulty Level				Conclusion
	(a_i)	Explanation	(b_i) 1	(b_i) 2	(b_i) 3 average	Explanation	
Sik_01	0,129	Good	-1,35	0,584	-0,383	Good	Good
Sik_02	0,848	Good	-0,036	0,281	0,123	Good	Good
Sik_03	0,982	Good	-1,545	2,091	0,273	Good	Good
Sik_04	1,274	Good	0,452	0,624	0,538	Good	Good
Sik_05	1,213	Good	-1,712	0,897	-0,408	Good	Good
Sik_06	1,635	Good	-0,717	-0,459	-0,588	Good	Good
Sik_07	0,813	Good	-0,92	1,296	0,188	Good	Good
Sik_08	-0,071	Bad/Low	-2,456	4,299	0,922	Good	Corrected
Sik_09	1,577	Good	0,78	1,766	1,273	Good	Good
Sik_10	1,586	Good	-0,115	1,093	0,489	Good	Good
Sik_11	0,15	Good	-2,719	2,905	0,093	Good	Good
Sik_12	1,057	Good	-0,18	2,091	0,956	Good	Good
Sik_13	0,579	Good	0,14	-1,264	-0,562	Good	Good
Sik_14	0,053	Good	-0,469	0,027	-0,221	Good	Good
Sik_15	0,37	Good	-1,666	1,924	0,129	Good	Good

The table above is the parameter estimation result. The table shows good discrimination (a_i) except for item Sik_08. Item Sik_08 has discrimination index which is -0.071 and the discrimination includes in bad category or low category.

The difficulty level of the item (bi) shows 4 categories of difficulty level, i.e., (bi) 1 and (bi) 2. The difficulty level of the item in each category shows different results, but if it is viewed based on the average of difficulty level shows that all items have difficulty level which is in good category.

Based on the table, it can be found that all items have good difficulty level and discrimination, except for the item Sik_08. The item Sik_08 has bad discrimination but has a good difficulty level. Based on this case, the item Sik_08 will be corrected.

The data on the table x shows that all items (15 items) have discrimination index in the range of 0 to 2. Except for the item Sik_08, the item has discrimination index -0.071 or it does not meet the criteria of good discrimination index. According to the estimation result of difficulty level, obtained the information that 15 items of attitude assessment have difficulty index which is in good category since they are in the range of -2 to 2 in all 15 items.

Referring to the analysis estimation result of discrimination and difficulty level, it can be conceived that all items have good discrimination and difficulty level, except for item Sik_08. Although the item Sik_08 has bad discrimination, but it has good difficulty level therefore it will be conducted the correction on item Sik_08.

Moreover, the researcher will reveal the information function of attitude assessment instrument which is in the form of questionnaire. The representation of the information function of assessment instrument can be observed through the graph of information function and error measurement on figure 6 below.

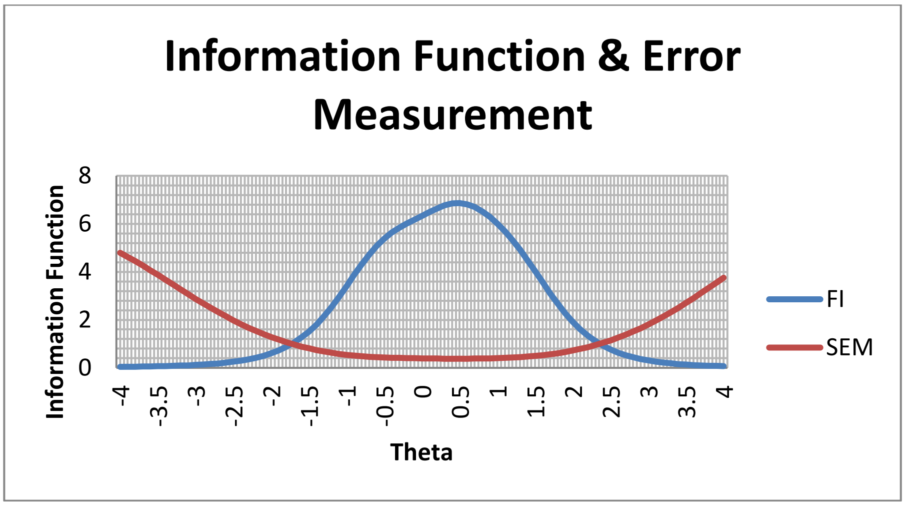

Figure 6. Graph of Information Function and Error Measurement of Attitude Assessment Instrument

Figure 6 shows that the attitude assessment instrument acquire the maximum information which is 6.863 on theta which is about 0.5 and have error measurement which is 0.500. the value of information function obtained is on the range of skill -1.75 to +2.25. this case indicates that the attitude assessment instrument is suitable for students with the ability -1.75 to +2.25 or students with low to high ability.

DISCUSSION

The initial prototype of MP-KPMO which has been studied qualitatively by experts includes several components of the model which will be immediately corrected by the researcher. Suggestions and inputs from experts include reducing the components of the model instrument specification and model syntax, as the model instrument is already included in the model guide and the syntax is included in the model component. After the prototype model is improved, then the document is submitted back to the experts as validators. Experts check whether the improvements have been appropriate with the suggestions and inputs given. After being declared appropriate, then the experts provide a quantitative assessment.

The improvement of the assessment instrument of mathematical reasoning skill is based on the suggestion and input given at the stage of expert validation and readability tests. At the validation stage, input and suggestion include instrument construct improvement, grids, item questions, and scoring guidelines. Improvement of the instrument items include changing the context and the size of the questions for Basic Competency 3.9., i.e., a rectangular cat cage 24 cm2 in area is replaced with a rectangular city park with an area of 48 m2. Changing the location of the image from a position on the right edge to a position in the middle, the problem in Basic Competency 4.7 is area and circumference of a flat shape on the skill dimension. Stimulus in the form of donuts was replaced with Bakpia to make it more contextual. In the attitude dimension instrument, the input and suggestion given by the experts include constructs, instrument grids, and statement items. The Improvements of instrument grid are in the form of using appropriate terms, which are familiar with students. Meanwhile, the improvement of items includes item number 1 regarding interest and item number 13 regarding belief.

The knowledge, skill, and attitude assessment instrument that have been improved were then handed back to the experts to be checked whether they are appropriate with the suggestion or input that had been given. The experts provide a quantitative assessment to obtain content validity. After the assessment instrument meets the content validity criteria, readability tests are then carried out by academics, practitioners, teachers, and students. The second improvement was made to the assessment instrument based on input or suggestion from academics, practitioners, teachers, and students. Suggestion or input given, especially regarding the language in the formulation of questions, both in the dimension of knowledge and skill need to be simplified and communicated so that they are easily understood by students.

The next improvement was carried out during instrument testing, including the skill assessment instrument on the knowledge dimension and skill dimension. Referring to the parameter estimation result of the knowledge assessment instrument item, it is found out that item 2C or item number 2 for the indicator 'making or investigating allegations' in the Basic Competency of data presentation has good difficulty level, but it has bad discrimination. Furthermore, these items are corrected with an emphasis on discrimination. Meanwhile, the result of the estimation of discrimination and difficulty level of the skill assessment instrument item show that items 1A, 3A, and 11C have low difficulty level, but have good discrimination.

Therefore, improvements are made for point 1A or item number 1 and item number 3 in the Basic Competency 4.7 which involves indicators analyzing information/data and making or investigating allegations, as well as item 11C or item number 3 in the Basic Competency 4.11 involving the indicator 'making or investigating allegations'. Improvements are made especially on the stem or subject matter.

CONCLUSION

The analysis result of knowledge, skill and attitude assessment instrument item that have been conducted by GPCM/2PL approach are as follow.

1. Knowledge assessment instrument, all items of the knowledge assessment instrument are in good category with discrimination index 0.422 2.401 and difficulty index -1.31 to 0.924 on the logit scale. There is one item that should be corrected since it has good difficulty level but has low discrimination. The assessment instrument provides accurate information in the range of skill -2.5 - 1.7 with the highest information on skill which is about -0.8.
2. The skill assessment instrument, in general all items skill assessment instrument is in good category with discrimination index 0.561 – 1.554 and difficulty index -2.335 – 1.808. There are 3 questions that should be corrected since they have low difficulty level but have good discrimination. The skill assessment instrument provides accurate information on the skill range -3.2 – 2.6 with the highest information on the skill which is about 1.3.
3. In general, all items of the attitude assessment instrument are in good category with a discrimination index -0.071 – 1.635 and difficulty index -0.588 – 1.273. There are 3 questions that should be corrected because they have low difficulty level but have good discrimination. The attitude assessment instrument provides accurate information in the skill range of -1.75 – 2.5 with the highest information on skill which is about 0.5.

References

1. Badrun Kartowagiran & Amat Jaedun. (2016). . Model assessment autentik untuk menilai hasil belajar siswa Sekolah Menengah Pertama (SMP): Implementasi assessment autentik di SMP. *Jurnal Penelitian dan Evaluasi Pendidikan*, 20 (2), 131 – 141.
2. BSNP. (2010). *ParadigmaPendidikan Nasional Abad XXI*. [Online]. Tersedia: http://www.bsnpindonesia.org/id/wpcontent/uploads/2012/04/LaporanBSNP-2010.pdf diakses pada tanggal 11 Maret 2015
3. DeMars, C. (2010) *Item Response Theory: Understanding Statistics Measuremen*t. Oxford: Oxford University Press.
4. Ebel, R.L. & Frisbie, D.A. (1991). *Essentials of educational measurement*. Fifth edition. New Delhi: Prentice Hall of India.
5. Hadi, S. (2013). *Pengembangan Computerized Adaptive Test Berbasis Web*. Yogyakarta: Aswaja Pressindo.
6. Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional module on educational measurement: issues and practice. *Comparison of classical test theory and item response theory and their applications to test development*, 12(3), 38-47.
7. Harlanu, M., Suryanto, A., Ramadhan, S., & Wuryandini, E. (2023). Construct validity of the instrument of digital skill literacy. Cakrawala Pendidikan, 42(3), 781–790. Scopus. https://doi.org/10.21831/cp.v42i3.59703
8. Kismoyo, C. P., Kartowagiran, B., Suyanto, S., & Ramadhan, S. (2023). Analysis of Continuity of Care Research Developments Based on the Scopus Database 20 20-2023: A Bibliometric Study. International Journal of Membrane Science and Technology, 10(3), 164–173. Scopus. https://doi.org/10.15379/ijmst.v10i3.1499
9. Mullis, I.V.S., & Martin, M.O. (Eds.). (2014). *TIMSS advanced 2015 assessment frameworks*. Chestnut Hill, MA: Boston College.
10. Muraki, E. (1999). *New appoaches to measurement*. Dalam Masters, G.N. dan Keeves, J.P.(Eds). Advances in measurement in educational research and assesment. Amsterdam : Pergamon.
11. Muraki,E., & Bock, R.D. (1997). P*arscale 3: IRT based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software Inc.
12. NCTM. (2000). *Principle an Standars for School Mathematics*. Reston, VA: NCTM.
13. Nunnally, J.C.(1981). Psychometric theory (2nd ed). New York : Mc Grawhill, Inc.
14. Priatna, N. (2003). *Kemampuan penalaran dan pemahaman matematika siswa sekolah lanjutan tingkat pertama negeri di Kota Bandung*. Disertasi Doktor pada PPs IKIP Bandung Press: Tidak Diterbitkan
15. Retnawati, H. (2014). *Membuktikan validitas instrumen dalam pengukuran*. Diambil pada tanggal 8 Juli 2015 dari http://www.evaluation-edu.com.
16. Retnawati, H. (2015). *Validitas, Reliabilitas & Karakteristik Butir*. (Panduan untuk Peneliti, Mahasiswa, dan Psikometrian). Yogyakarta: Parama Publishing.

17. Rivo Pujo Yudo. (2016). Pengembangan instrumen otentik unjuk kerja materi bangun ruang di sekolah dasar kota Cirebon. *EduMa*, 5(2), 2-11.
18. Sri Wardhani dan Rumiati .(2011). *Instrumen Penilaian Hasil Belajar Matematika SMP: Belajar dari PISA dan TIMSS*. Yogyakarta: Kementerian Pendidikan Nasional, Pusat Pengembangan dan Pemberdayaan Pendidik dan Tenaga Kependidikan.
19. Sri Wardhani. (2006). *Model pembelajaran Matematika dengan pendekatan berbasis masalah*. Yogyakarta: PPPG
20. Stiggins, R. J. (1987). *The design and development of performance assessments*. *Educational Measurement: Issues and Practice, 6*, 33-42.
21. Sujaya, AAGR, Suarni, Ni Ketut, dan Candiasa, I Made. (2013). *Pengaruh model pembelajaran asesmen autentik terhadap hasil belajar Matematika dengan kovariabel motivasi berprestasi (eksperimen pada siswa kelas V SD negeri 1 Gianyar)*. e-Journal Program Pascasarjana Universitas Pendidikan Ganesha, Program Studi Penelitian dan Evaluasi Pendidikan, 3, (1 – 12).
22. Umar, J. (1999). Item Banking. Dalam Masters, G.N. dan Keeves, J.P. (Ed). *Advances in Measurement in Educational Research and Assessment (*pp.207-218). New York: Pergamon.
23. Wahyudin. (2008). *Pembelajaran dan model pembelajaran (pelengkap untuk meningkatkan pedagogis para guru dan calon-calon profesional*). Bandung: UPI
24. Zuroidah, N., Kumaidi, Hadi, S., Kusaeri, & Ramadhan, S. (2024). The Joint Model of Two-Parameter Logistic and Response Time Model for Computer-Based Tests. International Journal of Engineering Trends and Technology, 72(1), 117–129. Scopus. https://doi.org/10.14445/22315381/IJETT-V72I1P112