Rasch Model: Analyzing the Items Quality of Mathematics Higher-Order Thinking Skill Instrument

Muhamad Syahidul Qirom, Elah Nurlaelah
Universitas Pendidikan Indonesia

E-mail: msqirom@upi.edu

Abstract: This research aims to analyze a mathematics HOTS instrument using the Rasch Model to provide a better measurement instrument for mathematics. The result showed, based on Rasch Model analysis, 1) from 20 items, there are only ten items that can be used to measure HOTS of students in mathematics with the difficulty of 3 items being hard and seven items being easy; 2) the reliabilities of the instrument are moderate; 3) the instrument only can divide students into two groups of level ability; and 4) there is no significant differential item functioning (DIF) bias detected on the instrument.

Keyword: Rasch Model, HOTS, Mathematics, Item Quality, Assessment and Evaluation

INTRODUCTION

HOTS, also known as higher-order thinking skills, are processes of transfer, critical thinking, and problem-solving (Brookhart, 2010). Transfer means that Higher-Order Thinking Skill is not just about recalling something that has been learned but how the lesson becomes meaningful. Critical thinking means that students can provide a consideration or a critical reason for a problem through Higher-Order Thinking Skills. Meanwhile, problem-solving means that Higher-Order Thinking Skills prepare students to identify and solve problems in everyday life. According to Butterworth and Thwaites (2013), higher-order thinking skills include the ability to think critically (critical thinking), reason, and reflect (Reflection). Meanwhile, in Bloom's Revised Taxonomy, HOTS refers to the ability to analyze, evaluate, and create (Anderson et al., 2001). That shows the importance of higher-order thinking skills for students because it can help them develop critical thinking and problem-solving skills, which are part of 21st-century skills (Chiruguru, 2020).

It will undoubtedly positively impact students to prepare themselves to face various changes and challenges in the 21st century by improving 21st-century skills. So, teachers must train students' higher-order thinking skills to support 21st-century skills. Therefore, a HOTS assessment instrument, in the form of questions, is needed to help students to train their higher-order thinking skills. In addition, the existence of the instruments can help teachers evaluate the level of higher-order thinking skills that students have achieved.

A good instrument certainly has several aspects that must be met, such as validity and reliability. Assessing a tool's validity involves determining if it accurately measures what it is designed to measure (Petra & Aziz, 2020). While reliability refers to the consistency of measurement values obtained under the same conditions with the same instrument in repeated measurements (Sürücü & Maslakci, 2020). In addition, other aspects that are no less important are the difficulty index and discriminant power. It is necessary to analyze the question items to find out the achievement of these aspects in a question instrument. Item analysis is essential for optimizing items used in subsequent tests and minimizing factually inaccurate items in an instrument (Quaigrain & Arhin, 2017).

This study is a previous follow-up research on developing HOTS questions in mathematics subjects (Qirom, Sridana, & Prayitno, 2020). The analysis was carried out in previous studies with Classical Test Theory (CTT). However, using CTT to analyze the instrument has some things that could be improved. The main weakness was that no matter what attribute level an individual has, CTT estimates measurement precision as equal (Jabrayilov, Emons, & Sijtsma, 2016). It made generalizing its estimators difficult, especially when examinees were diverse in their abilities. (Bichi, Talib, Embong, & Salleh, 2019). As a result, if the

test were given to the weak ability of students, the difficulty and discrimination power of questions would be low. However, it would be high if it was given to students with solid abilities (Haw, Sharif, & K. Han, 2022). So, in this study, another analysis method will be used, namely item response theory (IRT), by using the Rasch Model.

Unlike CTT, in IRT, a fundamental assumption was that the probability of a student answering an item depended on the item's difficulty and the examinees' ability (Tavakol & Dennick, 2013). In addition, the item responses can be discrete or continuous and assessed dichotomously or polychotomous; item score categories can be sorted or unsorted; test performance can be based on one or more abilities; and there are numerous ways (i.e., models) to determine the relationship between item responses and the underlying ability or abilities (Hambleton & Jones, 2005). It makes applying the IRT more flexible than CTT for all data conditions. The IRT model method is often used to develop and assess scales and measures, especially in the education field (Yang Kao). It is because the IRT can predict a person's score based on latent abilities or traits and establish a relationship between a person's item performance and a set of traits underlying the item's performance (Hambleton, Swaminathan, & Rogers, 1991). Therefore, IRT does not only provide information regarding the item/instrument of the assessment but also the person or examinee's behavior. For example, when we use CTT for the analyzing instruments, it only provides us with item difficulty; there is no information regarding whether, for some examinees, the item is complicated or which item the difficulty of an item is over students' ability. Knowing the kind of information is essential while developing instruments for assessment. It can guide the evaluation of the instrument to provide better, more valid, and more reliable results.

As part of IRT, Rasch Model measurement is a method of analyzing response data in which both the items' test and examinee are integrated into a mathematical model to predict how each examinee will answer each question in the test (Karlin & Karlin, 2018). Rasch analysis is intended to increase the accuracy of researchers in instrument construction, instrument quality monitoring, and responder performance calculation (Boone, 2016). One of the most significant components of Rasch measurement is the ability to explain the meaning of student actions and group actions using the context of the instrument's items; if a group of students increases from pre to post, the researcher can explain why (Boone, 2016). Rasch Model also can help to generate better interpretations about what an individual can or cannot do in qualitative descriptions is its advantage (Zamora-Araya, Smith-Castro, Montero-Rojas, & Moreira-Mora, 2018). So that by using the Rasch model, an instrument that provides a more accurate interpretation of students' HOTS in mathematics can be obtained. The Rasch model has three assumptions (Michalos, 2014b):

1) The number of items correctly answered by the examinee adequately measures the test taker's aptitude.
2) The likelihood of the test taker correctly answering the question increases with his competence.
3) There is no item interaction; responding correctly to one item does not affect the test taker's ability to reply correctly to other things.

Research about analyzing the quality of HOTS instruments, especially for mathematics subjects, was conducted before (such as How, Zulnaidi, & Rahim, 2023; Yudha, 2023). However, How et al (2023) only analyze the HOTS question for only one sub-topic, namely the quadratic equation. In addition, they also use classical test theory (CTT) for analyzing the instrument. Meanwhile, Yudha (2023) developed the mathematics HOTS instrument for mathematics topics in seventh grade. Nevertheless, he used the Rasch model only for analyzing the instrument's reliability.

Therefore, considering the advantages of IRT to CTT, this study will analyze mathematics HOTS instruments using the Rasch Model. Also, the analysis will emphasize almost all aspects of proper assessment instrument criteria: validity, reliability, discrimination power, difficulty index, and DIF. It is also rare to find research about the analysis or development of instruments using the Rasch Model. As a result, this research also provides a better understanding of analyzing the assessment instrument using the Rasch Model. Thus, the study uses the Rasch Model measurement to analyze HOTS items in mathematics for the junior high school level.

METHOD

Research Design

This research is a descriptive quantitative design. Descriptive research gathers information about variables without modifying or manipulating

any variables, either individuals or conditions (Gay, Mills, & Airasian, 2012). The purpose of this type of research is to describe a phenomenon naturally (Gay et al., 2012; Nassaji, 2015) from a collection and analysis of numerical data (Mertler, 2015).

Data Collection

The data was collected from 25 students in grade nine in a public junior high school in Mataram, West Nusa Tenggara, Indonesia. The instrument is a higher-order thinking skills instrument in mathematics. The instrument consisted of 20 multiple-choice questions with four options. Students must choose only one answer among the options given for each question. For each correct answer, they will get one point, and no point for wrong answer. The questions develop based on Revised Taxonomy Bloom, which includes analysis (C4), evaluation (C5), and creation (C6) levels. The mathematics topics for the question also vary. They are numbers, algebra, measurement and geometry, statistics and probability. The distribution of cognitive level for each topic and subtopic was presented in Table 1. One question for each sub-topic.

Data Analysis

Five paramount concerns regarding item quality were analyzed in this research: validity, reliability, difficulty index, discriminant, and Differential Item Functioning (DIF) bias. Those aspects were analyzed using Rasch Model measurement with the Ministeps application.

Table 1. Distribution of Topic and Cognitive Level

Cognitive Level	Sub-topic	Topic
	Linear Equation with One Variable	Algebra
	Arithmetic Series	Number
	Ratio	Number
C4 (Analysis)	Number Pattern	Number
	Function	Algebra
	Circle	Measurement & Geometry
	Line and Angle	Measurement & Geometry
	Integer	Number
	Social Arithmetic	Number
	Fraction	Number
	Algebra Form	Algebra
	Quadrilateral	Measurement & Geometry
C5 (Evaluation)	Central Tendency	Statistics & Probability
	Data Presentation	Statistics & Probability
	System of Linear Equation with Two Variables	Algebra
	Set	Algebra
	3D-Shape with Flat Surface	Measurement & Geometry
C6 (Creation)	Line Equation	Algebra
	Triangle	Measurement & Geometry
	Probability of A Chance	Statistics & Probability

a. Validity Criteria

To see whether each item is suitable or feasible to be used to measure HOTS, it can be seen in the MNSQ Outfit, ZSTD Outfit, and PTMEASURE Corr. Sections (Table 2). In this analysis, if one criterion does not meet the item fit value, then the question needs to be corrected. If the question does not meet two or more item fit values, the question must be discarded.

Table 2. Item Fit

Criterion	Item Fit Value
Outfit mean square (MNSQ)	$0.5 < MNSQ < 1.5$
Outfit Z-standard (ZSTD)	$-2.0 < ZSTD < +2.0$
Point measure correlation	$PTMeasure\ Corr. > 0.4$

(Bond & Fox, 2015; Boone, Staver, & Yale, 2014)

b. Reliability Criteria

In this study, the reliability analysis refers to Rasch Reliabilities, namely Person and Item Reliability. The reliability criteria will follow Table 3 (Hinton, McMurray, & Brownlow, 2014).

Table 3. Reliability

Reliability Score	Category
≥ 0.75	High
$0.5 \leq Score < 0.75$	Moderate (Acceptance threshold)
< 0.5	Low

Table 4. Difficulty Index Category

Measure Value	Category
$Measure < -SD*$	Very Easy
$-SD \leq measure < 0$	Easy
$0 \leq measure < +SD$	Hard
$Measure \geq +SD$	Very Hard

*SD : Standard Deviation

c. *Discriminant Criteria*

To determine the discrimination power (D) of a person (student) or question item, the following formula can be used (Linacre, 2022):

$$D = \frac{(4 \times separation)+1}{3} \quad (1)$$

d. *Difficulty Index Criteria*

The Difficulty Index in the Rasch Model uses the following Table 4 (Karlimah, 2022). However, the category is flexible. It depends on the separation index of the item.

e. *DIF Bias Criteria*

DIF for each question item based on the Rasch Model will be determined based on two criteria (Linacre, 2022):

1) If |DIF Contrast| > 0.5, DIF is likely to occur.
2) If there is a possibility of DIF then the t value will be checked. If $|t| > 2$, DIF significantly occurs in a question item (5% significance level).

In this study DIF analysis was used to detect bias between male and female students' difficulty while answering the questions.

Table 5. Item Fit Result

Item (Question) Number	JMLE Measure	Outfit		PTMEASURE Corr.
		MNSQ	ZSTD	
16	1.86	0.90	0.19	0.24
9	1.37	2.10	1.45	-0.21
15	1.37	1.40	0.74	0.10
14	1.01	0.69	-0.43	0.38
17	0.44	0.88	-0.18	0.52
18	0.35	1.07	0.30	0.21
20	0.21	1.00	0.10	0.35
7	0.11	0.79	-0.60	0.53
3	-0.01	1.29	1.01	0.13
19	-0.01	1.03	0.20	0.48
8	-0.22	0.89	-0.38	0.44
10	-0.22	1.01	0.11	0.29
2	-0.41	0.91	-0.33	0.46
4	-0.41	1.36	1.54	0.10
12	-0.41	0.95	-0.16	0.49
5	-0.60	1.14	0.75	0.28
11	-0.72	0.76	-1.22	0.58
13	-0.98	0.76	-1.29	0.59
1	-1.36	0.98	-0.03	0.35
6	-1.36	0.97	-0.06	0.48

Table 6. Summary of Item Fit for Each Question

Question	Outfit (MNSQ)	Outfit (ZSTD)	PT Measure Corr.	Decision
2, 6, 7, 8, 11, 12, 13, 17, 19	Satisfied	Satisfied	Satisfied	Retained
1, 3, 4, 5, 10, 14, 15, 16, 18, 20	Satisfied	Satisfied	Unsatisfied	Repaired
9	Unsatisfied	Satisfied	Unsatisfied	Discarded

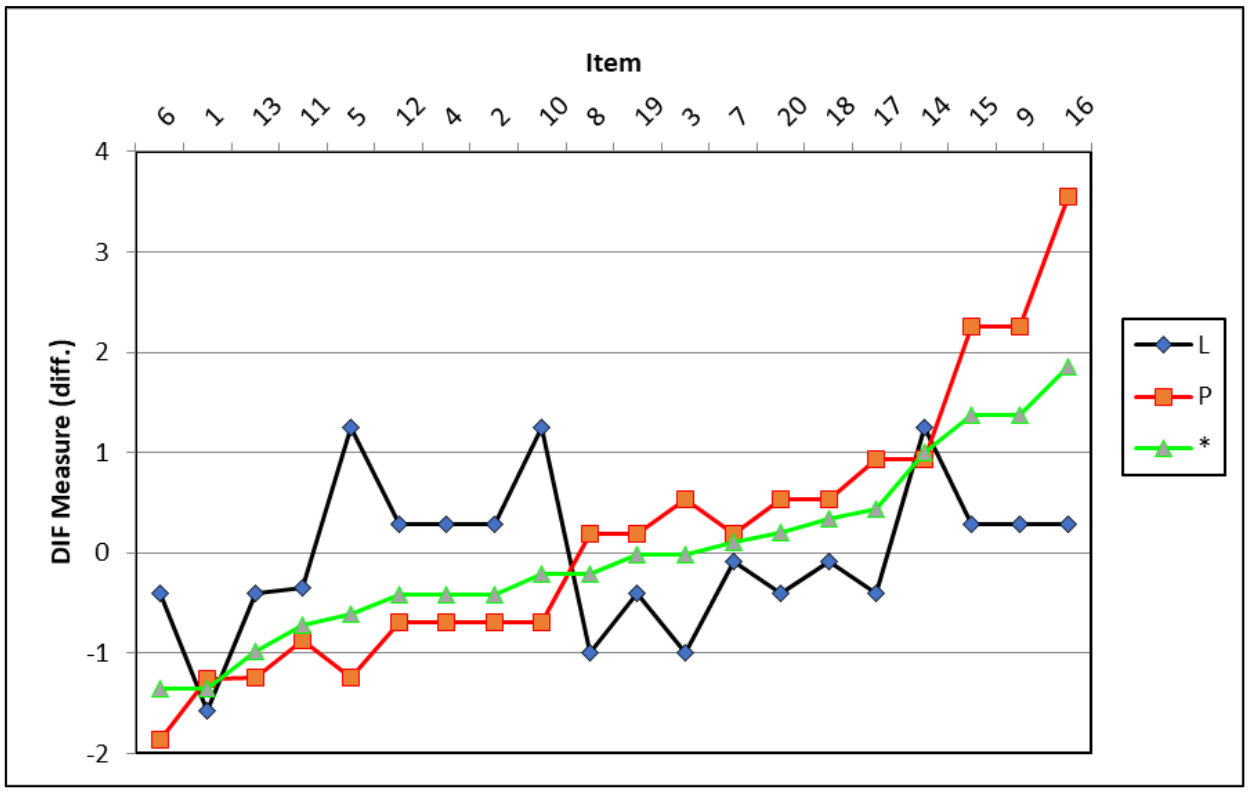

Figure 1. DIF Measure of Each Question

RESULT AND DISCUSSION

Validity

Table 6 shows the questions that fulfill the item fit for the validity test. Based on Table 6, nine questions must be retained since they fulfill all the criteria. Ten questions are considered to be repaired. One question is considered discarded since it only fulfills one item fit criteria.

Reliability

Table 7 shows the reliability of the instrument. Both person and item reliability are in moderate category. Furthermore, the separation and personal reliability values are 1.14 and 0.57 (moderate), respectively, indicating that there is a Low Separation Person because the separation value is < 2 and person reliability is < 0.8. The separation and item reliability values are 1.32 and 0.63 (moderate), respectively, indicating that there is Low Separation Item because the separation value is < 3 and item reliability is < 0.9.

Table 7. Summary of Rasch Reliabilities

Test	Score	Category
Person Separation	1.14	
Person reliability	0.57	Moderate
Item Separation	1.32	
Item reliability	0.62	Moderate

Discriminant Power

With a personal separation value of 1.14, using formula (1), a value of D = 1.85≈ 2 is obtained; this indicates that the instrument can only classify students into two groups of ability levels. With an item separation value of 1.32, using formula (1), a value of D = 2.09 ≈ 2 is obtained, indicating that the instrument can only be divided into two difficulty levels.

Index difficulty

Table 8 shows the difficulty category of each question/item. Based on *item separation value,* the difficulty level of questions can only be separated into two categories: hard and easy. If the question's JMLE Measure Value (see Table 5) is more than or equal to 0, then the question is hard. Meanwhile, if the JMLE Measure of the question is less than 0, then the question is easy.

Table 8. Difficulty of Question

Question	Difficulty
1, 2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 19	Easy
7, 9, 14, 15, 16, 17, 18, 20	Hard

DIF bias

Figure 1 shows the DIF bias of male and female students for each question. The green, blue, and red lines represent the overall measure without DIF in terms of the question's difficulty, the

difficulty of an item (question) for male students, and the difficulty of an item (question) for female students, respectively. For example, look at question number 6, use the green point as the reference point, and see that the blue point value is higher than the green point. Meanwhile, the red point value is under the green point. It indicates that question number 6 seems more difficult for male than female students. However, it is hard to determine whether the DIF bias happened in the questions only from the figure above. Therefore, we need to follow the guideline aforementioned to check whether the DIF bias is significant. Table 9 shows that based on DIF Contrast, DIF bias does not appear only in three questions, while the rest do. However, the DIF on the questions shows insignificant value.

Table 9. DIF Bias Significance

Question	DIF Contrast	Significance	Decision
1, 7, 14	No DIF identified	-	All question free of DIF bias
2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20	DIF identified	Not Significance	

DISCUSSION

Based on the analysis results with the Rasch Model, a summary of the decision on questions based on item fit, whether they are retained, repaired, or discarded, can be seen in Table 6. Note that the Point Measure Correlation (PTMeasure Corr.) value for item number 9 is negative (Table 5). A negative Point Measure Correlation value indicates the opposite direction, where low-ability students can answer the question but not high-ability students. This kind of question is considered to be discarded. Other questions that did not meet the Point measure correlation value (PTMeasue Corr.) were considered for improvement. As for the questions that have met the three qualified items, the questions are suitable to measure students' higher-order thinking skills for mathematics subjects; in other words, these items have met the validity aspects based on the Rasch Model.

The reliability of the instrument, either person or item, is in moderate category. A Reliability person in a moderate category means that the instrument moderately discriminates the students into enough levels. Meanwhile Moderate item reliability means that the sample (students) is moderately enough to precisely locate the item's difficulty. Since the Rasch reliability is moderate, then the instrument is reliable. The discrimination power can be seen through the separation value and person reliability (Table 7). Because the separation value and personal reliability show low separation, this indicates that the instrument is not too sensitive to be used to distinguish or categorize students' abilities (Linacre, 2022). In line with the low separation of persons, low separation also occurred in the question items. It indicates that the sample was not large enough to confirm the difficulty level of the instrument (Linacre, 2022). In other words, this instrument is difficult to differentiate the difficulty level. However, when viewed in more detail using Formula 1, it is obtained that the instrument can only divide student ability into two groups if viewed from the aspect of student ability. Likewise, if viewed from the aspect of the difficulty level of the question, the instrument can only divide the level of ability into two levels too.

In addition, for checking the level of difficulty of the question can be seen from the JMLE Measure column (Table 5). Pay attention to the Entry Number column. The column shows the order of the questions from the most difficult (top) to the easiest (bottom). Because the item separation analysis results only identified two categories or difficulty levels, in this study, the question level is divided into two levels of difficulty, namely questions with easy and hard categories.

.

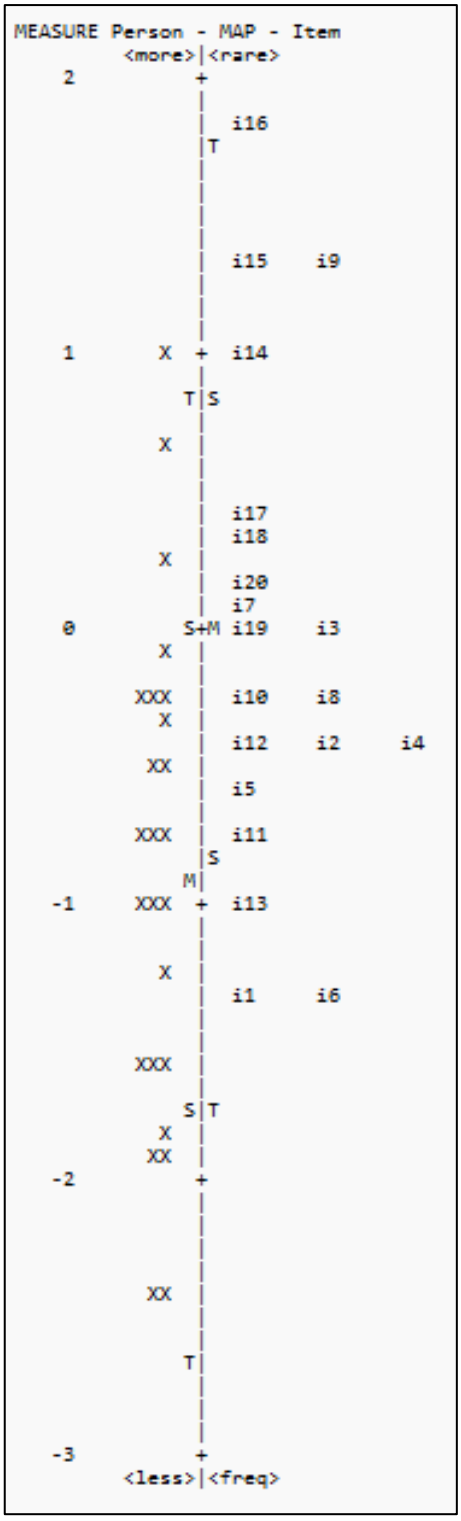

Figure 2. Person-Item Wright Map

Furthermore, Figure 2 shows the Wright map to compare person (students) ability and the item (questions) difficulty. The Wright map consists of two vertical histograms. The left side displays students, while the right side displays questions. The left side of the map depicts the distribution of the students' measured ability, with the most competent students at the top and the least one at the bottom. Meanwhile, the right side of the Wright map arranges the questions based on its difficulty, with the most challenging question at the top and the least difficult at the bottom. In Figure 2, the mean (M) and two standard deviations (SD) points (S = one SD and T = two SD) for measured students' ability on the left side and on the right side for the questions. The Wright map in Figure 2 also demonstrates that the students' mean (M) ability is approximately one standard deviation (S) below the questions' mean (M) difficulty. It indicates that the instrument's difficulty exceeds the students' ability; the questions seem complicated. Figure 2 also shows that there are eight students (see "X" on the left side) that have the ability under all the difficulty levels of questions, even the easiest questions (i1 and i6). Those students probably cannot answer almost all the questions correctly. Furthermore, students cannot pass the difficulty level of questions 16 (i16) and 15 (i15). It means all students' abilities are under the difficulty level of those questions. It indicates that almost all students answer both questions wrongly.

The Rasch Model also can help to determine DIF bias on an item. DIF refers to an item that causes bias when it is administered to subjects from different groups with the same quality, but in response to a specific item, they have varying probabilities (Scott et al., 2010). For example, an item that is worked on by groups of men and women of equal ability has a higher level of difficulty (unfair) for one of the groups only (Michalos, 2014a). The item functions differently for different genders, even though both groups have equal abilities. Take a look at Figure 1, showing the level of DIF bias between the male group (blue (L)) and the female group (red (P)). The higher the gap point on the graph, the more complex the item is for that group. Figure 1 shows that the DIF measures distance between L and P is the greatest for item numbers 3, 5, 6, 8, 9, 10, 15, 16, and 17. It shows a considerable difference in the difficulty level between men and women on these items. In questions 5, 6, and 10, male students tended to have difficulty answering these questions, but for women, these questions were elementary. Whereas for questions 3, 8, 9, 15, 16, and 17, female students tended to have more difficulty answering them than male students. However, the primary determinant to decide whether DIF occurs significantly or not in a question item is by checking the contrast value and the t-test value of the question. Table 9 shows that all question items did not detect significant DIF. Therefore, all of the items are free of DIF.

Based on the analysis of the mathematics HOTS instrument with the Rasch Model, the question items that can be used to measure students' HOTS level are question items number 2, 6, 7, 8, 11, 12, 13, 17, and 19. Other question items are considered to be increased in difficulty level, considering that most of the question difficulty indexes are in the easy category. As for question number 9 will be discarded. However, since question item number 14 has PTMeasure Corr. value 0.38, which is close to the threshold value of 0.4. Also, according to Boone et al (2014), it is sufficient to look at the MNSQ Outfit to find out whether the instrument is feasible. Thus, the item will also be included as the feasibility question for assessing the HOTS level of students. So that ten questions can be used directly to measure students' HOTS in mathematics subjects (numbers 2, 6, 7, 8, 11, 12, 13, 14, 17, and 19), one question is discarded.

REFERENCES

Anderson, L. W., Krathwhol, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., … Wittrock, M. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objective*. New York: Addison Wesley Longman.

Bichi, A. A., Talib, R., Embong, R., & Salleh, S. salleh. (2019). Comparative Analysis of Classical Test Theory and Item Response Theory using Chemistry Test Data. *International Journal of Engineering and Advanced Technology*, *8*(5C), 1260–1266. https://doi.org/10.35940/ijeat.E1179.0585C1

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Third edition). New York: Routledge.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. Alexandria, Va: ASCD.

Butterworth, J., & Thwaites, G. (2013). *Thinking Skills: Critical Thinking and Problem Solving* (Secon Edition). Cambridge: Cambridge University Press.

Chiruguru, S. (2020). *The Essential Skills of 21st Century Classroom (4Cs)*. https://doi.org/10.13140/RG.2.2.36190.59201

Gay, L. R., Mills, G. E., & Airasian, P. W. (2012). *Educational research: Competencies for analysis and applications* (10th ed). Boston: Pearson.

Hambleton, R. K., & Jones, R. W. (2005). An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, Calif: Sage Publications.

Haw, L. H., Sharif, S. B., & K. Han, C. G. (2022). Analyzing the science achievement test: Perspective of classical test theory and Rasch analysis. *International Journal of Evaluation and Research in Education (IJERE)*, *11*(4), 1714. https://doi.org/10.11591/ijere.v11i4.22304

Hinton, P. R., McMurray, I., & Brownlow, C. (2014). *SPSS explained* (Second edition). London: Routledge.

How, R. P. T. K., Zulnaidi, H., & Rahim, S. S. B. A. (2023). Development of Higher-Order Thinking Skills test instrument on Quadratic Equation (HOTS-QE) for secondary school students. *Pegem Journal of Education and Instruction*, *13*(1). https://doi.org/10.47750/pegegog.13.01.41

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement*, *40*(8), 559–572. https://doi.org/10.1177/0146621616664046

Karlimah, K. (2022). How does Rasch modeling reveal difficulty and suitability level the fraction test question? *Jurnal Elemen*, *8*(1), 66–76. https://doi.org/10.29408/jel.v8i1.4170

Karlin, O., & Karlin, S. (2018). Making Better Tests with the Rasch Measurement Model. *InSight: A Journal of Scholarly Teaching*, *13*, 76–100. https://doi.org/10.46504/14201805ka

Linacre, J. M. (2022). A User's Guide to WINSTEP-MINISTEP: Rasch-model computer program. Retrieved from winsteps.com

Mertler, C. A. (2015). Quantitative Research Methods. In *Introduction to Educational Research* (First Edition). California: SAGE Publications.

Michalos, A. C. (Ed.). (2014a). *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5

Michalos, A. C. (Ed.). (2014b). Rasch Analysis. In *Encyclopedia of Quality of Life and Well-Being Research* (pp. 5393–5395). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-0753-5_2394

Nassaji, H. (2015). Qualitative and descriptive research: Data type versus data analysis. *Language Teaching*

Research, *19*(2), 129–132. https://doi.org/10.1177/1362168815572747

Petra, T. Z. H. T., & Aziz, M. J. A. (2020). Investigating reliability and validity of student performance assessment in Higher Education using Rasch Model. *Journal of Physics: Conference Series*, *1529*(4), 042088. https://doi.org/10.1088/1742-6596/1529/4/042088

Qirom, M. S., Sridana, N., & Prayitno, S. (2020). Pengembangan Soal Matematika Berbasis Higher Order Thinking Skills Pada Lingkup Materi Ujian Nasional Untuk Tingkatan Sekolah Menengah Pertama. *Jurnal Pijar Mipa*, *15*(5), 466–472. https://doi.org/10.29303/jpm.v15i5.2028

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., … Sprangers, M. A. (2010). Differential item functioning (DIF) analyzes of health-related quality of life instruments using logistic regression. *Health and Quality of Life Outcomes*, *8*(1), 81. https://doi.org/10.1186/1477-7525-8-81

Sürücü, L., & Maslakçi, A. (2020). Validity and reliability in quantitative research. *Business & Management Studies: An International Journal*, *8*(3), 2694–2726. https://doi.org/10.15295/bmij.v8i3.1540

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher*, *35*(1), e838–e848. https://doi.org/10.3109/0142159X.2012.737488

Yudha, R. P. (2023). Higher Order Thinking Skills (HOTS) Test Instrument: Validity and Reliability Analysis With The Rasch Model. *EduMa: Mathematics Education Learning And Teaching*, *12*(1), 21–38. http://dx.doi.org/10.24235/eduma.v12i1.9468

Zamora-Araya, J. A., Smith-Castro, V., Montero-Rojas, E., & Moreira-Mora, T. E. (2018). Advantages of the Rasch Model for Analysis and Interpretation of Attitudes: The Case of the Benevolent Sexism Subscale. *Revista Evaluar*, *18*(3). https://doi.org/10.35670/1667-4545.v18.n3.22201