THE ANALYSIS OF TEACHER-MADE TEST ITEMS IN MADRASAH IBTIDAIYAH EXAMINATIONS USING THE RASCH MODEL

M. Saufi Rahman
STAI Rakha Amuntai, Indonesia
Email: msaufirahman87@gmail.com

Norhayati
STAI Rakha Amuntai, Indonesia
Email**:** norhayati77@gmail.com

ABSTRACT
This study aimed to analyze the quality of teacher-made tests in the Madrasah Examination for the Fikih of sixth-grade at Madrasah Ibtidaiyah Negeri 2 Hulu Sungai Tengah. The study employed quantitative evaluation method involving 125 students as research participants. The instrument analyzed consisted of 20 items. Data analysis was conducted using the Winsteps program to determine the fit of the test items to the Rasch model and to identify the characteristics of item difficulty levels. The results showed that all test items met the model fit criteria based on the MNSQ, ZSTD and Pt-Measure Corr values. These findings indicate that the instrument was able to measure students' abilities appropriately. The item difficulty analysis revealed that 18 items were categorized as good difficulty levels because they measure values within the range of -2.0 to +2.0. The distribution of items was still dominated by easy and moderate questions, indicating that the measurement of students' HOTs not been optimal. This study concludes that teacher-made tests still require improvement in terms of difficulty level distribution and the development of HOTs, to provide important implications for teachers to improving the quality of learning through the development of instruments were valid, reliable, objective, and reflective.
Keywords*: Item Test Analysis, Teacher-made Test, Rasch Model, Assessment of Madrasah*

How to Cite Rahman, M. S., & Norhayati, N. (2026). The Analysis of Teacher-Made Test Items in Madrasah Ibtidaiyah Examinations Using the Rasch Model. *Al-Mubtadi: Jurnal Pendidikan Guru Madrasah Ibtidaiyah*, *3*(2), 510–523. https://doi.org/10.58988/almubtadi.v3i2.1163

INTRODUCTION

Improving the quality of education cannot be separated from the quality of the learning evaluation system used in madrasah. The learning process and the evaluation of learning outcomes are holistic relationship in improving the quality of learning. Learning evaluation is an important component of the education system because it functions to determine the level of achievement of learning objectives, the effectiveness of the learning process, and the quality of students' learning outcomes, as well as serving as a basis for educational decision-making (Jonnasen, 2013); (Rahman et al., 2025); & (Utami et al., 2025). In assessment and learning evaluation activities, instruments are required to measure the level of learning achievement. High-quality evaluation instruments must fulfill the principles of validity, reliability, objectivity, practicality, and the ability to accurately distinguish students' competency levels (Mardapi, 2017); (Retnawati, 2016); (Rahman, 2025); & (Juwita et al., 2026). Therefore, the quality of teacher-made test items becomes an important factor in determining the quality of learning evaluation outcomes in educational institutions.

One form of assessment and learning evaluation activity is the madrasah examination at the Madrasah Ibtidaiyah (MI) level. The purpose of the madrasah examination in Madrasah Ibtidaiyah is to serve as an important activity in measuring students' competency achievement after participating in the learning process (Wardah et al., 2022). Madrasah examinations hold a strategic position as they can be used to measure the standards of students' competency achievement at the end of the elementary education phase. Madrasah examination standards not only function as a tool for assessing learning outcomes, but also serve as indicators of the quality of instruction implemented by teachers throughout the educational process in a reflective and informative manner (Wiyogo et al., 2025).

Current phenomena indicate that the quality of teacher-made tests still faces many problems, particularly in the aspects of item construction and empirical item analysis. Based on observations of teacher-made test items, many madrasah examination questions are still categorized as too easy or too difficult, resulting in a disproportionate distribution of difficulty levels. This phenomenon occurs because many teachers still have limited understanding of

how to develop good test items that are reflective and informative regarding student learning outcomes, causing the quality of many test instruments to remain unidentified (Wijayama et al., 2024); (Hajar et al., 2026). As a result, evaluation activities conducted tend to be superficial and unclear. Therefore, item analysis is needed to improve the quality of test items through revision and the elimination of unfit items. In the modern era, item analysis is not merely numerical in nature but can also be reflective and informative for students (Magdalena, 2025); (Jafar et al., 2026). In Classically that item analysis refers to the examination of test items through information obtained from students' responses, which functions to improve the quality of test items.

Therefore, as a strategy to improve the quality of test items in the 21st century, teachers are expected to design questions that can measure critical thinking skills and enhance students' analytical abilities. The ability to assess students' competencies has become an essential aspect of modern learning evaluation in order to develop students' analytical, critical, creative, and problem-solving skills (Handayani & Wulandari, 2021) & (Maslihah et al., 2025). Teacher-made test items in elementary schools and madrasahs are still predominantly focused on measuring lower-order cognitive skills, such as remembering and understanding. To address this issue, item analysis is needed as a strategic effort for teachers to identify the quality of the tests they develop.

To determine the quality of teacher-made test items in the madrasah examination, the author employed analysis using the Rasch Model approach. The Rasch Model is considered a modern test analysis model compared to Classical Test Theory. The Rasch Model is capable of generating an item map that demonstrates whether the test developed can accommodate various student characteristics in the classroom(Haryanto, 2020). The analysis of teacher-made tests is important because it can provide empirical information regarding the quality of madrasah examination instruments used in Grade VI Madrasah Ibtidaiyah. Item analysis can identify the quality of each test item based on validity, reliability, level of difficulty, discriminating power, and distractor effectiveness, so that poorly constructed items can be revised before being reused.

Based on the introduction above, research on the analysis of teacher-made tests in Grade VI Madrasah Ibtidaiyah examinations needs to be conducted in order to determine the quality of the evaluation instruments used in madrasahs. This study is expected to provide theoretical contributions to the development of learning evaluation and to serve as a reflective reference for teachers and madrasahs in improving the quality of test construction that is valid, reliable, and aligned with the demands of modern educational assessment that is both informative and reflective.

METHOD

This study employed a quantitative descriptive evaluation approach. The data analyzed consisted of 20 teacher-made test items used in the madrasah examination for the Fikih subject at Madrasah Ibtidaiyah Negeri 2 Hulu Sungai Tengah. The sample in this study involved 125 students selected through purposive sampling techniques. The content validity of the teacher-made test instrument was examined through expert validation involving five experts using Aiken's criteria calculation technique. The data analysis technique applied was the Rasch Model analysis using the Winsteps program. Item analysis using the Rasch Model provides information regarding item fit. If the analysis indicates that an item is not fit, it suggests the presence of errors within the item, which can be useful for teachers in improving the quality of the test items. In Rasch Model analysis, item quality is determined based on item fit criteria. The criteria used to determine item fit were: (1) acceptable Outfit Mean Square (MNSQ) values of $0.5 < MNSQ < 1.5$; (2) acceptable Outfit Z-Standard (ZSTD) values of $-2.0 < ZSTD < +2.0$; and (3) acceptable Point Measure Correlation (Pt Mean Corr) values of $0.4 <$ Pt Measure Corr < 0.85. In the Winsteps analysis, an ideal mean square value is 1.0, while the standardized value (Z-standardized value) is 0.0.

RESULTS AND DISCUSSION

RESULT

The results of the analysis of the 20 madrasah examination items for the Fikih subject at Madrasah Ibtidaiyah Negeri 2 Hulu Sungai Tengah using the Rasch Model approach indicate that the

quality of the teacher-made test instrument is still in the moderate category and requires improvement in several measurement aspects. Based on the reliability analysis, the instrument demonstrated fairly good consistency in measuring students' abilities. However, several test items were found to be misfitting with the Rasch Model. The misfit of these items indicates that some questions were not yet able to measure students' competencies accurately and consistently.

These findings indicate that test construction still faces challenges in the aspects of item development and the alignment between learning indicators and the forms of questions designed by teachers. The critical criteria for item fit in the Winsteps program include an Outfit MNSQ value ranging from 0.5 to 1.5, an Outfit ZSTD value ranging from -2 to +2, and a positive Pt-Measure Correlation value. If these criteria are met, the item can be considered fit or compatible with the model. An item is categorized as fit if it fulfills at least one of the three criteria. The following are the results of the item fit analysis using the Winsteps program.

Tabel 1. Item Fit of the Context Assessment Instrument

Outfit MNSQ	Outfit ZSTD	PT-Measure Corr	Item
.99	.1	.58	Item_09
.99	.0	.50	Item_15
1.08	.4	.46	Item_20
.97	-.1	.47	Item_05
.98	.0	.49	Item_16
1.08	.4	.39	Item_10
1.14	.7	.36	Item_11
.73	-1.2	.56	Item_17
1.95	2.5	.23	Item_12
1.69	2.0	.19	Item_14
.86	-.2	.36	Item_04
.84	-.3	.37	Item_08
.70	-.7	.44	Item_18
.68	-.7	.46	Item_19
.77	-.2	.29	Item_03
.95	.1	.25	Item_07
.64	-.1	.25	Item_02
.64	-.1	.25	Item_13
.64	-.1	.25	Item_06
.42	-.1	.23	Item_01

Based on the table.1 above, concluded that the results of the analysis of the madrasah examination instrument for the Fikih subject using the Rasch Model approach indicate that the quality of the test items generally meets the criteria for model fit. The analysis was conducted on 20 test items involving 125 Grade VI students at Madrasah Ibtidaiyah Negeri 2 Hulu Sungai Tengah. Item fit testing was carried out through the values of Outfit Mean Square (MNSQ), Outfit Z-Standard (ZSTD), and Point Measure Correlation (Pt-Measure Corr). Based on the results presented in Table 1, all test items were declared suitable or fit to the Rasch Model because they fulfilled at least one of the feasibility criteria, namely an Outfit MNSQ value within the range of 0.5–1.5, an Outfit ZSTD value within the range of -2 to +2, and a positive Pt-Measure Corr value.

The test items with the highest level of fit were Item_15 and Item_09, each having an Outfit MNSQ value of 0.99 and Pt-Measure Corr values of 0.58 and 0.50, respectively. These values indicate that both items were able to measure the construct of students' abilities consistently and in accordance with the expectations of the model. On the other hand, several items approached the threshold of model misfit, such as Item_12 with an Outfit MNSQ value of 1.95 and an Outfit ZSTD value of 2.5, as well as Item_14 with an Outfit MNSQ value of 1.69 and an Outfit ZSTD value of 2.0. Nevertheless, both items were still considered acceptable because their Pt-Measure Corr values remained positive. This indicates that, in general, all items were still able to contribute to measuring students' abilities.

These findings indicate that the instrument used possesses fairly good empirical quality. This condition supports the view, Bond & Fox who stated that the Rasch Model is capable of objectively detecting the fit of test items to the measurement construct through probabilistic analysis (Bond & Fox, 2013). Instruments that fit the model indicate that students' response patterns align with the expectations of the measurement model, thereby making the evaluation results more valid and reliable. Furthermore, Retnawati explained that model fit analysis is important to ensure that each item truly measures the same competency and does not create measurement bias (Retnawati, 2016).

The results of this study also indicate that most items had positive Pt-Measure Corr values ranging from 0.19 to 0.58. These

positive correlation values suggest that all test items were able to distinguish between students with high and low abilities. However, several items such as Item_12, Item_14, and Item_01 showed relatively low correlation values and therefore require attention during the item revision process. Items with low correlation values are usually caused by unclear item wording, distractors that do not function optimally, or levels of difficulty that are too extreme.

Furthermore, item characteristic analysis was conducted through the level of item difficulty based on the measure values in the Rasch Model. Based on the analysis results presented in Table 2, the difficulty levels of the test items varied from very difficult to very easy categories. The item with the highest difficulty level was Item_15 with a measure value of 2.18, while the lowest difficulty level was found in Item_01 with a measure value of -2.39. According to the Rasch Model criteria, an item is considered good if it falls within the range of -2.0 to +2.0. Therefore, Item_15 and Item_01 were categorized as less satisfactory because they were outside the ideal difficulty range.

Item_15 was categorized as too difficult because it only obtained a total score of 39 out of 125 respondents. This indicates that most students experienced difficulties in answering the item. An excessively high level of difficulty may be caused by the complexity of the material, the use of language that is difficult to understand, or questions that require higher-order thinking skills that students are not yet accustomed to solving. In contrast, Item_01 was categorized as too easy because almost all students were able to answer it correctly. Items that are too easy tend to be less effective in distinguishing students' abilities because they do not provide sufficient challenge.

Tabel 2. Output of the Analysis of Item Characteristics of the Context Assessment Instrument

Number	Total Score	Total Count	Measure	Model S.E
15	39	125	2.18	.52
9	63	125	1.27	.46
20	63	125	1.27	.46
5	68	125	1.06	.45
16	68	125	1.06	.45
10	71	125	.86	.45
11	71	125	.86	.45

Number	Total Score	Total Count	Measure	Model S.E
17	75	125	.66	.45
12	85	125	.24	.47
14	85	125	.24	.47
4	90	125	-.22	.50
8	90	125	-.22	.50
18	90	125	-.22	.50
19	90	125	-.22	.50
3	93	125	-.78	.57
7	93	125	-.78	.57
2	96	125	-1.62	.75
6	96	125	-1.62	.75
13	96	125	-1.62	.75
1	98	125	-2.39	1.03

These findings are consistent with the opinion Retnawani who stated that the quality of an evaluation instrument is highly determined by the suitability of test items to the competency construct being measured (Retnawati, 2016). Items that do not fit the model indicate the possibility of ambiguous wording, mismatch of indicators, or distractors that do not function optimally. In addition, Mardapi emphasized that a good instrument must be able to produce objective and consistent data so that the evaluation results truly reflect students' abilities (Mardapi, 2017). In the context of this study, several items were found to be too easy and too difficult, making them unable to distinguish students' abilities proportionally. The difficulty level analysis showed that the distribution of item difficulty was not yet balanced. Most items were categorized as easy, while several others were classified as very difficult. This condition indicates that the construction of test items had not yet considered an ideal proportion of difficulty levels. In fact, a balanced distribution of item difficulty is very important for comprehensively measuring students' abilities. Items that are too easy tend to measure only memorization skills, while items that are too difficult may hinder students from demonstrating their competencies optimally.

DISCUSSION

The results of the analysis of the item characteristics of the context assessment instrument using the Rasch Model approach indicate that most test items have met the criteria for an appropriate

level of difficulty. Based on the analysis results, out of 20 test items, 18 items were within the ideal Rasch Model difficulty range of -2.0 to +2.0. These findings indicate that, in general, the instrument has been able to measure students' abilities proportionally and in accordance with the characteristics of the respondents' abilities. This is in line with the statement of Hayati et al, who explained that Rasch analysis is able to provide more objective estimates of item difficulty because it simultaneously considers the interaction between students' abilities and item characteristics (Hayati et al., 2024). Therefore, items that fall within the ideal range indicate that the instrument possesses fairly good measurement quality.

The results of the study indicate that Item 15 was the test item with the highest level of difficulty, having a measure value of 2.18. This value exceeds the ideal Rasch Model threshold, therefore the item was categorized as too difficult. The high level of difficulty indicates that only a small number of students were able to answer the item correctly. This condition may be caused by several factors, such as the complexity of the material, the use of language that is difficult to understand, or the demand for higher-order thinking skills that are not yet aligned with the average ability of the students. According to Mardapi and Hariyanto that items with excessively high difficulty levels tend to make the instrument less effective in measuring competencies and may reduce students' motivation in completing the test because the questions are perceived to be beyond their level of ability (Haryanto, 2020; Mardapi, 2017).

These findings also indicate that Item 15 had the lowest total score compared to the other items, obtaining only 39 points out of 125 respondents. This demonstrates that the item was not yet able to accommodate students' abilities evenly. In the context of learning evaluation, items that are too difficult need to be revised by considering the alignment of indicators, the level of language complexity, and the relevance of the material to the competencies that students have learned. Bond and Fox emphasized that the Rasch Model can help researchers identify overly difficult items so that teachers can revise the instrument before it is used again (Bond & Fox, 2013).

In contrast, Item 1 became the item with the lowest level of difficulty, with a measure value of -2.39. This value indicates that the

item was categorized as very easy because almost all students were able to answer it correctly. Items that are too easy are less effective for evaluation purposes because they are unable to distinguish between students with high and low abilities. This finding is consistent with the opinions of Wright & Nitko and Brookhart who stated that overly easy items tend to produce a homogeneous score distribution, thereby reducing the discriminating power of the instrument (Brookhart & Nitko, 2011; Wright, 2007). Item 1 obtained a total score of 98 out of 125 respondents, indicating that nearly all students were able to answer the question without significant difficulty.

However, most of the other test items were categorized as having a moderate level of difficulty. Item 9 and Item 20 had measure values of 1.27, while Item 5 and Item 16 had measure values of 1.06. These values indicate that the items possessed a fairly proportional level of difficulty and were able to provide challenges to students without excessively burdening them. According to Rahman that items with a moderate level of difficulty represent the most ideal type of test item because they are better able to distinguish students based on their levels of ability more accurately (Rahman et al., 2025). In addition, several items such as Item 4, Item 8, Item 18, and Item 19 had measure values of -0.22, indicating that the items were relatively easy but still within the ideal category. The presence of easy items in an evaluation instrument remains important to measure students' basic abilities and to provide opportunities for lower-ability students to demonstrate their competencies. Nevertheless, the distribution of items within an evaluation instrument should remain balanced among easy, moderate, and difficult items so that the evaluation results are able to comprehensively represent students' abilities.

The results of this study indicate that the distribution of item difficulty levels was still dominated by easy and moderate questions rather than difficult ones. This condition suggests that the evaluation instrument still predominantly measures lower-order thinking skills (LOTS). This finding is consistent with Widyaningsih et al. who stated that teachers still experience difficulties in developing Higher Order Thinking Skills (HOTS)-based questions, particularly in the aspects of analysis, evaluation, and problem-solvin (Widyaningsih et

al., 2021). In fact, the demands of 21st-century learning require teachers to design evaluation instruments that are capable of developing students' critical, creative, collaborative, and communicative thinking skills in order to generate learning information that is reflective and informative.

Maslihah et al. explained that modern learning evaluation not only functions to measure learning outcomes, but also serves as a means of training students' higher-order thinking skills (Maslihah et al., 2025). Therefore, the development of HOTS-based questions needs to be carried out continuously so that learning evaluation can support the strengthening of 21st-century competencies. In Madrasah Ibtidaiyah, the development of high-quality evaluation instruments is also important because evaluation results are not only used to determine students' final scores, but also serve as reflective material for improving the quality of learning.

The use of the Rasch Model in this study provided more comprehensive information regarding the quality of the evaluation instrument. The Rasch Model not only identified the level of item difficulty, but also helped determine which items needed to be revised or retained. Polat et al. explained that the advantage of the Rasch Model compared to Classical Test Theory lies in its ability to produce measurements that are more objective, linear, and independent of sample characteristics (Polat et al., 2022). Therefore, the results of the analysis can be used as a basis for continuously improving the quality of evaluation instruments.

CONCLUSION

The madrasah examination instrument for the Fikih subject at Madrasah Ibtidaiyah Negeri 2 Hulu Sungai Tengah generally demonstrated fairly good quality based on the Rasch Model analysis. The item fit analysis showed that all test items met the model fit criteria because they possessed Outfit MNSQ, Outfit ZSTD, and Pt-Measure Corr values that were consistent with Rasch measurement standards. These findings indicate that the instrument was able to measure students' abilities consistently, objectively, and in accordance with the competency constructs being assessed. The results of the study also revealed that the evaluation **instrument** was still dominated by items with easy and moderate characteristics,

indicating that the measurement of students' higher-order thinking skills had not yet been optimal. This condition suggests that the development of Higher Order Thinking Skills (HOTs) based questions still needs to be improved so that learning evaluation does not merely assess memorization and comprehension abilities, but also analytical, critical, and problem-solving skills in accordance with the demands of 21st-century learning. Therefore, these analytical findings indicate that teachers need to enhance the use of critical-thinking-based stimuli in the learning process so that students' abilities in responding to teacher-made test items can improve more effectively.

REFERENCES

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Brookhart, S. M., & Nitko, A. J. (2011). *Strategies for constructing assessments of higher order thinking skills*. Prentice-Hall, Inc.

Hajar, A., Jamaluddin, J., Amir, A., Asmah, S., Rasyid, A. T., Yahya, R., & Asnidar, A. (2026). *Evaluasi Pembelajaran (Konsep, Teknik Penilaian, Analisis, dan Implementasi di Era Digital)*. CV Eureka Media Aksara.

Handayani, R., & Wulandari, D. (2021). Modern assessment dalam menyongsong pembelajaran abad 21 dan hambatan di negara berkembang. *Jurnal Pendidikan Edutama*, *8*(1), 13–22.

Haryanto, M. P. (2020). *Evaluasi pembelajaran (konsep dan manajemen)*. UNY Press.

Hayati, S., Hidayanti, S. W., & Rizki, A. (2024). Analisis Rasch Model: Mengukur Kualitas Butir Soal Tes Tashrif Lughawi. *Intelektualita: Journal of Education Sciences and Teacher Training*, *13*(1).

Jafar, M., Nur, S., & Firdaus, M. R. (2026). *Evaluasi pembelajaran peserta didik (EDISI 2)*. CV. Ruang Tentor.

Jonnasen, D. A. (2013). *Evaluating constructivistic learning. In Constructivism and the technology of instruction*. Routledge.

Juwita, R., Hasania, Y. A., & Azis, A. (2026). Penyusunan Instrumen Evaluasi Pembelajaran: Analisis Pembuatan Soal Pilihan Ganda dan Esai/Uraian. *Al-Ilmiya: Jurnal Pendidikan Islam*, *1*(4), 1837–1844.

Magdalena, I. (2025). *Implementasi Pembelajaran Abad 21 dan Bank Soal*. Fahmi Karya.

Mardapi, D. (2017). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Parama Publishing.

Maslihah, A., Aziroh, K. M. U., & Bashith, A. (2025). Strategi efektif dalam evaluasi penilaian pembelajaran berbasis HOTS untuk meningkatkan kompetensi kognitif siswa sekolah dasar. *Jurnal Ilmiah Pendidikan Citra Bakti*, *12*(1), 94–106.

Polat, M., Turhan, N. S., & Toraman, C. (2022). Comparison of classical test theory vs. Multi-facet Rasch theory in writing assessment. *Pegem Journal of Education and Instruction*, *12*(2), 213–225.

Rahman, M. S. (2025). Bank of english test based on google form: Critical thinking test development of the sumative final semester (sas) english at mts (islamic junior high school). *SKETCH JOURNAL: Journal of English Teaching, Literature and Linguistics*, *5*(1), 17–25.

Rahman, Riffani, & Mukarramah. (2025). Critical thinking test development of the sumative final semester (sas) english at mts (islamic junior high school) as bank of english test based on google form. *Journal of English Educational Study (JEES)*, *8*(2), 176–184.

Retnawati, H. (2016). *Validitas reliabilitas dan karakteristik butir*. Parama Publishing.

Utami, A. P., Balqis, Z., Apriani, L., Rahmadani, L., & Amanda, R. D. (2025). Peran Evaluasi Pembelajaran Dalam Upaya Peningkatan Kualitas Pendidikan Di Madrasah Ibtidaiyah. *Al-Hasib: Jurnal Manajemen Pendidikan Islam*, *2*(2), 434–446.

Wardah, D. M. M., Hadi, N., & Purwasih, J. H. G. (2022). Implementasi ujian madrasah di Madrasah Aliyah Miftahul Ulum Kalirejo Kabupaten Pasuruan. *Jurnal Integrasi Dan Harmoni Inovatif Ilmu-Ilmu Sosial (JIHI3S)*, *2*(2), 548–556.

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-Learning of Moodle LMS. *International Journal of Instruction*, *14*(4), 51–68.

Wijayama, B., Farda, U. J., Maulida, A. H., Fauziya, L., & Hardiyanti, S. (2024). *Asesmen Pembelajaran SD/MI Kurikulum Merdeka.* Cahya Ghani Recovery.

Wiyogo, A. P., Chaniago, A. S., Handayani, A., Nugroho, D. T., & Prianti, H. (2025). Evaluasi Pembelajaran Sebagai Sarana Refleksi dan Peserta Didik. *Edu Society: Jurnal Pendidikan, Ilmu Sosial Dan Pengabdian Kepada Masyarakat*, *5*(3), 1391–1397.

Wright, R. J. (2007). *Educational assessment: Tests and measurements in the age of accountability*. Sage Publications.