Volume 8 Number 2 (2026)
May-August 2026
Page: 457-473

E-ISSN: 2656-4491
P-ISSN: 2656-4548
DOI: 10.37680/scaffolding.v8i2.9191

SCAFFOLDING
Jurnal Pendidikan Islam dan Multikulturalisme

HIGHER ORDER THINKING-BASED FIQH EVALUATION INSTRUMENT FOR SENIOR HIGH SCHOOL FOR THE ANALYSIS OF STUDENTS' COMPLEX COGNITIVE ABILITIES IN ISLAMIC RELIGIOUS EDUCATION

Dwi Lestari[1], Risnita[2], Dedi Sastradika[3], Sri Ramdayeni Sakunti[4]
[1234]Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi; Indonesia
Correspondence E-mail; risnita@uinjambi.ac.id

Submitted: 08/02/2026 Revised: 31/03/2026 Accepted: 23/06/2026 Published: 01/07/2026

Abstract

The main objective of this study is to develop an HOTS instrument in the Islamic Religious Education subject, fiqh material, in the form of 40 multiple-choice questions. This study is a Research and Development (R&D) study using the 4D model, limited to the Development stage. The experimental subjects in this study were 313 12th-grade high school students in Jambi City. Data were collected through expert validation and instrument trials. Data analysis was carried out using Aiken's V index and the Rasch model item response theory using QUEST software. The Aiken's V results reached 0.95, which is classified as very valid. Meanwhile, the results of the Rasch model analysis showed an item reliability value of 0.96 and a person reliability value of 0.90. 75% of the questions developed were included in the medium difficulty category, and there were 36 items that were declared appropriate and 4 items that were declared inappropriate. Based on these results, it can be concluded that the HOTS instrument for fiqh material meets the criteria of being valid, reliable, and suitable for use, although there are still several items that need to be revised.

Keywords Assessment, Fiqh, HOTS, Rasch Model.

INTRODUCTION

Islamic religious education in the contemporary virtual generation needs to be incorporated with spiritual values and better-order wondering skills. This objectives to form students who are able to adapt to technological developments while still having a strong spiritual foundation in accordance with Islamic teachings (Tolchah & Mu'ammar, 2019). However, in reality, learning in the subject of Islamic religious education still faces various problems, including limited competence in the use of the era and digital media, which makes it difficult for instructors to integrate generations into the PAI studying procedure (Ismail et al., 2025; Ju'subaidi et al., 2025), so that the learning methods used are still conventional and traditional in nature, causing them to be less interesting and not in accordance with the challenges of the times (Tatang Hidayat, 2019; Wiliyanti et al., 2023). The breadth of virtual interaction requires guidance and material filtering approaches to reduce deviations from Sharia values (Rashed et al., 2025). However, teachers' readiness and training in integrating Islamic values effectively are still very lacking (Amalia et al., 2025).

Teaching Islamic Religious Education, especially fiqh, has many problems. Fiqh material, which is abstract and complex in nature, as well as the very large accumulation of fiqh literature, demands deep understanding and critical analysis (Aldossari, 2021). However, the learning that is carried out still uses conventional learning methods that emphasize memorization and theoretical knowledge, so it is less stimulating to students' critical and analytical thinking skills when facing real-life problems (Haryani et al., 2021) and also difficulties in explaining the basic principles of fiqh contextually (Yıldırım & Gülengül, 2025). Students also have difficulty understanding Arabic terms used in fiqh material, especially in discussions of inheritance law (faraid) and muamalah, as well as in understanding the meaning and context of arguments from the Qur'an and Hadith. Many students experience confusion when explaining basic concepts again, and in understanding legal sources directly from classical books (Syahidan & Mukminin, 2024). And the lack of deep understanding of fiqh concepts, as well as difficulties in understanding and applying fiqh rules in real life, causes the minimal ability of students to solve fiqh problems creatively and critically (Tuli & Munirah, 2022).

Various problems have negatively impacted the development of students' higher-order thinking skills, which are lacking in Islamic Religious Education (PAI), particularly in Islamic jurisprudence (fiqh). Previous research has found that students' low higher-order thinking skills are related to teachers' skill and competence in integrating Higher-Order Thinking Skills (HOTS) into

the curriculum and learning activities (Karwadi et al., 2024). Furthermore, teachers generally have poor skills in generating HOTS questions. This is evident in the large number of lower-order thinking skill questions in various trimesters and final semester exams prepared by teachers (Mawangir, 2022).

The results of previous research are in line with observations conducted in several high schools in Jambi, which show that the evaluation process of Islamic religious education learning, especially fiqh material, is still dominated by questions that measure the ability to remember and understand. Most of the questions used still focus on aspects of understanding, definition, mentioning, memorizing, and re-explaining the material that has been studied, for example, questions mentioning the definition of marriage, mentioning the conditions of marriage, and its pillars. In addition, the instruments used have not been fully able to measure students' abilities in analyzing, broadcasting, and solving problems related to daily life based on Islamic legal sources. Such situations really highlight the urgent need for instruments of HOTS-based assessment in the education field, as well as in Islamic Religious Education, especially regarding Material fiqh. Therefore, various previous studies have been conducted as efforts to expand assessment contraptions that might be able to assess students' higher-order thinking abilities.

The creation of assessment tools for Islamic spiritual education is primarily based on in advance studies, and fiqh material has been widely carried out. Research by Mawangir (2022) Constructed a Higher Order Thinking Skills (HOTS) instrument in the form of 21 multiple-choice questions, which cover the topic of fiqh for grade XI senior high school. The results also showed that the scale is valid and reliable. Research by Sa'idah & Isnaini (2020) also developed 15 HOTS questions for the fiqh subject for grade XI Madrasah Aliyah. This instrument was found to be reliable in assessing students' higher-order thinking skills. Meanwhile, Firdaus (2024) created a more comprehensive fiqh measurement tool, consisting of multiple-choice questions, essays, and assessments of skills and attitudes based on religious moderation, with results that meet good validity and reliability standards. Research by Thowiyyah & Hidayatulloh (2024) study shows that the construction of 40 HOTS-based multiple-choice questions is in line with testing students higher order thinkings skill. In contrast to these studies, Sutiyono (2022) created 30 multiple-choice questions for the fiqh difficulty in Madrasah Aliyah, however, the results showed that the developed instrument had not yet reached sufficiently good measurement tool standards.

In general, previous studies referring to the development of test instruments in Islamic Religious Education, especially for fiqh material, displayed a wide variance. Some studies have succeeded in creating instruments that have fairly good validity and reliability criteria, but there are studies that still have shortcomings, especially due to the limited number of respondents and the analysis methods used. These studies often use traditional types of analysis with small sample sizes and are therefore not able to make in-depth descriptions of the items and student abilities. Therefore, it is necessary to develop HOTS test instruments with a larger number of respondents and analyzed using a modern measurement approach.

In response to those problems, this study developed Higher Order Thinking Skills (HOTS) instruments for fiqh materials by using 4D's development model, which focused on the development phase. This research uses a fairly large sample, namely more than 300 senior high school students, and applies the Item Response Theory (IRT) approach in the form of the Rasch model with the help of the QUEST software. The purpose of this research is to develop HOTS test instruments in the subject of Islamic Religious Education, fiqh material, through the 4D development model that is empirically tested using the Rasch model approach, so that it is feasible to be used as an evaluation tool for students' higher-order thinking skills.

METHOD

The technique used in this research is the research and development (R&D) technique using the 4D model (define, design, develop, disseminate), which is limited to the development stage. This model was developed by Sivasailam Thiagarajan, Dorothy S. Semmel, dan Melvyn I. Semmel pada tahun 1974 (Sihombing, 2024). The stages in the 4D model are as follows:

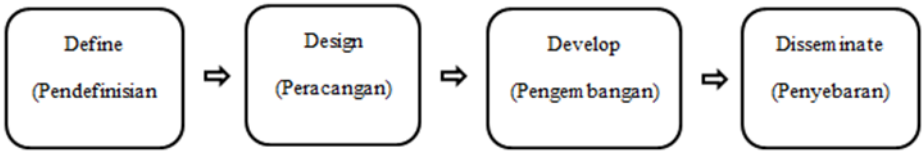

Figure 1. 4D Development Model

Primary and secondary sources were used to collect data in this study. Primary data were obtained through direct observation and the use of a HOTS-based multiple-choice reasoning test instrument from twelfth-grade high school students using Google Forms. Observations were conducted to determine learning conditions and the need for HOTS instrument development, while

the test aimed to measure students' higher-order thinking skills related to the research material. Conversely, books and scientific journals were among the supporting secondary data used to provide a theoretical framework for data analysis. The data source in this study was twelfth-grade high school students because they were deemed suitable for developing the HOTS instrument. Therefore, the data and data sources used in this study are expected to contribute to a comprehensive understanding and sufficiently support the achievement of this research objective.

The population used in this research is all twelfth-grade senior high school students in Jambi City who take the subject of Islamic Religious Education, fiqh material. The sampling technique in this study uses a purposive sampling technique, namely, a method of determining samples by taking into account certain considerations or criteria (Sugiyono, 2021). In this technique, the sample was taken with the criteria of twelfth-grade high school students, considering that the instrument used was a test designed according to the level of ability and material at that level. The sample used in this study consisted of 313 twelfth-grade students at State Senior High School 9, Jambi City. Three PAI lecturers, one PAI teacher, and one peer as instrument validators. Item validation involved three PAI lecturers, one PAI teacher, and one peer as instrument validators, with the assessed aspects being material suitability, suitability of the HOTS level, and language suitability using a scale of 4. The developed test instrument consists of 40 multiple-choice questions as follows.

Table 1. Question Instrument

Material	Sub-Material	Question Numbers According to Hots Level			Total Questions
		C4	C5	C6	
Fiqih	Khutbah, Tabligh, and Dakwah	9,11,14,17, 37, 38	2, 3, 19, 21, 23, 25	27, 28, 31, 32, 33	17
	Marriage	8, 12, 15, 18	7, 22, 26	1, 6, 29	10
	Inheritance	10, 13, 16	35, 39	4, 30	7
	Ijtihad	5, 40	20, 24	34, 36	6
Total Questions					40

The results of the validation sheets that have been filled in by several experts, namely three PAI lecturers, one PAI teacher, and one peer, using a scale of 4, will be analyzed using Aiken's V formula (Kuswanto, 2019) to determine the content validity of the developed instrument. If the validation test result is $0.81 < V \leq 1.0$, then the validity of the item is very good; if the validation test result is $0.41 < V \leq 0.80$, then the item is good; and if the analysis result is ≤ 0.4, then the item is very poor (Ahmad et al., 2024). After that, an item try-out is conducted, then the results of the item try-

out are analyzed using the Rasch model with the help of QUEST software. Rasch model analysis is used to measure various aspects related to the attributes or constructs of a measurement instrument, such as validity and reliability, item fit, item difficulty level, and respondent characteristics (Mufrihah, 2025). Aiken's V formula was chosen in this study because it involved a reasonable and valid approach to expert judgment to measure the content validity of an instrument. Aiken's V measures how appropriately the developed items reflect the actual indicators being measured (to a greater extent or degree) for higher instrument validity (Wahyuni, 2023). Simultaneously, a Rasch model was used to perform a deeper examination of item characteristics and respondent abilities. The Rasch model was chosen for its ability to provide information regarding item difficulty, fit, and overall reliability of the instrument. Applying the Rasch model help transforms ordinal data into interval data, improving the accuracy and objectivity of analytical results (Anggraini et al., 2023).

FINDINGS AND DISCUSSION

Findings

Results of HOTS Instrument Development

Define Stage: During the define phase, a review of literature from various journals was conducted via content analysis based on problems and needs in learning for the Islamic Religious Education of fiqh material. The results of this phase reveal that the examinations in Islamic Religious Education, such as those conducted in the assessment of fiqh, consist of questions that are more towards lower-order thinking skills (LOTS) rather than higher-order thinking skills (HOTS). The prevalence of this type, it prevents students from diving more deeply into fiqh, as they are not trained to think critically and analytically. It is therefore pertinent to develop an instrument of measurement in Islamic Religious Education that can capture the expression of high-order thinking skills on students' understanding and applying level concerning fiqh material.

Design Stage: At this designing phase, the HOTS instrument was developed to be consistent with the HOTS indicators, especially on analysis (C4), evaluation (C5), and creation (C6). This design also includes the preparation of a HOTS item blueprint that is adjusted to the Learning Outcomes (CP) in this subject. Furthermore, the item blueprint will be developed into questions at the development stage.

The develop Stage is the stage of constructing the questions that have been designed at the Design stage. At this stage, the questions are reviewed and validated by experts. In this research,

there are five validators, namely three lecturers, one PAI teacher, and one peer. They validated 40 items based on several criteria, such as language quality, suitability of verses or Hadith with the questions, the problems posed, and the correlation between questions with HOTS level. The evaluation was based on a 4-point Likert scale, and the content validity of each item was measured according to the Aiken V formula. Below is a table containing the average Aiken V results by all validators.

Table 2. Average Aiken V Results

Aspect	Description	Classification
Number of Items	40	-
Average Aiken V	0.95	Very Good

Source: Results of calculating the average Aiken's V using Microsoft Excel x

The results of Aiken's V analysis on the developed instrument were very good. All items, 100%, obtained Aiken's V coefficients in the very good category with an average value of 0.95. The high value indicates the quality of the substance and the suitability of the content of the developed instrument. To determine the feasibility of the items based on Aiken's V criteria, namely $0.81 < V \leq 1.00$ categorized as very good, $0.41 < V \leq 0.80$ categorized as good, and $V \leq 0.40$ categorized as very poor (Ahmad et al., 2024). Based on these criteria, the average Aiken's V value of 0.95 indicates that the HOTS instrument has very good content validity.

The high Aiken's V value indicates strong agreement among experts regarding the content validity of this HOTS instrument in fiqh material. Almost perfect agreement on the 40 items provides a strong methodological basis that this instrument truly represents the material domain well. Since no items fall into the low category, this instrument is considered content-valid and ready to proceed to the empirical testing stage to evaluate its construct validity and reliability.

Results of Rasch Model Analysis

The instrument that has been declared valid based on Aiken's V analysis was then tried out on 313 twelfth-grade senior high school students. After the item try-out was completed, the next stage was to conduct further analysis using the Rasch Model with the help of the QUEST program. This Rasch model analysis can identify items that do not fit (misfit) the model, measure the difficulty level of each item in the instrument (Mufrihah, 2025; Zeileis, 2025), and provide reliability indices for items and respondents. These indices show the extent to which the instrument consistently measures what it is supposed to measure. High reliability values indicate that the instrument is

reliable (Mufrihah, 2025; Ridwan et al., 2023; Zeileis, 2025). The Rasch Model also provides measures of respondents' abilities, allowing the identification of patterns in responses that may indicate inconsistency or misfit. This helps in understanding how well respondents engage with the instrument and whether their responses can be relied upon (Mufrihah, 2025; Yang et al., 2011; Zeileis, 2025). This model helps develop a strong and valid HOTS instrument to measure higher-order cognitive skills. This ensures that the tool used can measure students' abilities well and helps improve the quality of education.

Reliability Results

Reliability in the Rasch model has the function of showing the precision and stability of estimates at both the item and respondent levels (Caronni et al., 2023). The QUEST program calculates this reliability as part of the instrument quality assessment. The following presents the reliability results obtained from the analysis using the Quest program:

Table 3. Reliability

Aspect	Description
Summary of Item Estimates	0.96
Summary of Case Estimates	0.90

Source: Results of analysis using the Quest Program x

Based on Table 3, the reliability of the item estimates obtained is 0.96. This reliability value is included in the exceptional category because the item's reliability value exceeds 0.94 (Lestari et al., 2022). This shows that the items are stable, consistent, and fit the Rasch model, so that the estimation of difficulty levels can be trusted (Kamis et al., 2013). The person reliability value of 0.90 indicates consistent measurement, the ability to distinguish students' abilities accurately, and provides the same ability estimates for comparable sample groups. Overall, the HOTS instrument for fiqh material has an item reliability of 0.96 and a person reliability of 0.90, which confirms that the items are appropriate and feasible to be used as a valid and reliable measurement tool.

Item Fit Estimation

INFIT MNSQ is a key statistic used to evaluate the internal consistency of test or questionnaire items. This method helps identify items that do not fit the model and indicates potential problems in those items (Ariffin et al., 2010). According to Setyawarno (2017) in (Hanna & Retnawati, 2022; Rahayu & Putranta, 2025), the criteria for INFIT MNSQ values are >1.33 (not relevant to the Rasch model), 0.77–1.33 (relevant to the Rasch model), and < 0.77 (not relevant to the

Rasch model).

Table 4. Recapitulation of Item Fit

Item	INFIT MNSQ	Item	INFIT MNSQ	Item	INFIT MNSQ	Item	INFIT MNSQ
Item 1	1.55	Item 11	1.00	Item 21	0.69	Item 31	0.96
Item 2	0.93	Item 12	0.91	Item 22	0.86	Item 32	0.95
Item 3	1.70	Item 13	1.69	Item 23	1.62	Item 33	1.04
Item 4	0.92	Item 14	0.92	Item 24	0.69	Item 34	1.01
Item 5	0.82	Item 15	0.90	Item 25	0.77	Item 35	0.79
Item 6	1.00	Item 16	0.98	Item 26	0.88	Item 36	0.96
Item 7	1.14	Item 17	0.94	Item 27	0.87	Item 37	0.85
Item 8	0.93	Item 18	0.83	Item 28	0.69	Item 38	0.97
Item 9	1.05	Item 19	1.15	Item 29	1.06	Item 39	0.81
Item 10	0.93	Item 20	1.11	Item 30	0.75	Item 40	1.00

Source: Results of analysis using the Quest Program x

The results of the item fit analysis in Table 4 show that there are eight items, namely items 1, 3, 13, 21, 23, 24, 28, and 30, that have INFIT MNSQ values outside the range of 0.77 to 1.33, so they do not fit the Rasch Model (Hanna & Retnawati, 2022). Meanwhile, the other 32 instrument items are declared valid and fit the model. The conclusion regarding item validity is supported by the item fit map, which shows the position of each item relative to the fit boundary line.

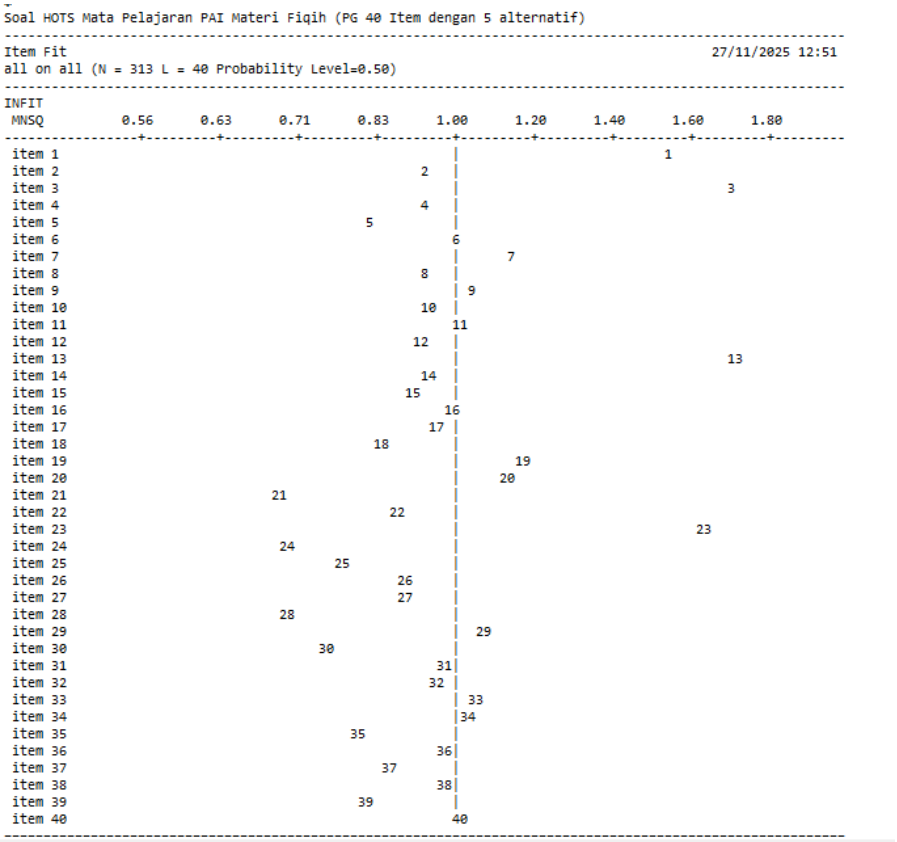

Figure 2. Fit Map Model Rasch

Based on Figure 2, it can be seen that there are several items that are outside the range of 0.77–1.33, namely items number 1, 3, 13, 21, 23, 24, 28, and 30, which indicates that these items do not fit the Rasch model. Thus, of the 40 HOTS instruments for fiqh material that were developed, there are 80% of items, totaling 32 items, that are considered relevant or fit the model, while the remaining 20%, totaling 8 items, are considered not to fit the model.

Difficulty Level

Analysis of item difficulty levels is very important to evaluate the suitability of items in measuring the targeted abilities. The results of the item difficulty analysis using the Quest program can be seen in the following figure.

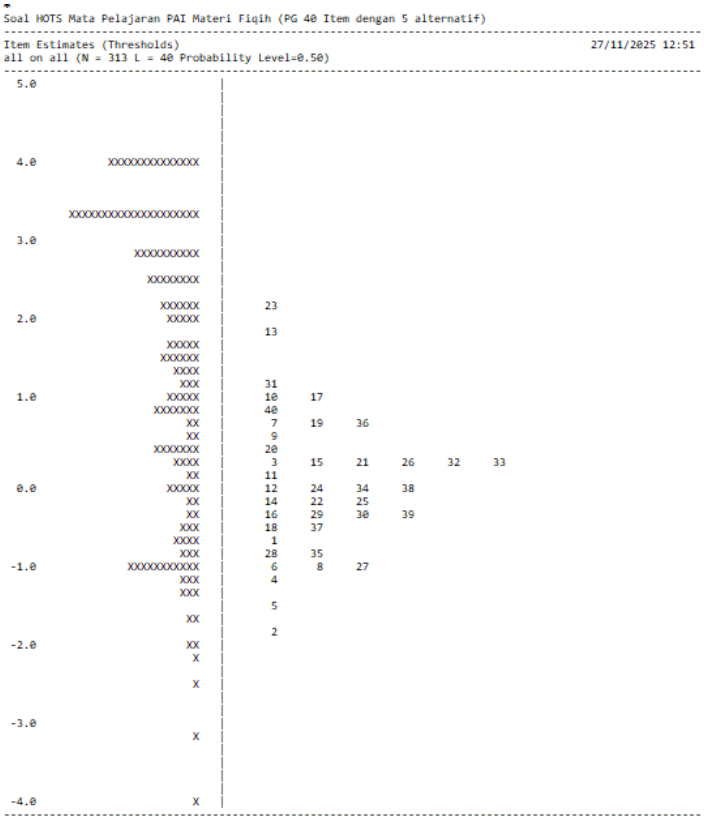

Figure 3. Item Difficulty Level

Figure 3 shows the distribution of difficulty level values of the developed HOTS items. From this figure, it can be identified which items are difficult, moderate, or easy according to the item difficulty criteria. The criteria for item difficulty levels are $b > 2$ (Very difficult), $1 < b \leq 2$ (Difficult), $-1 < b \leq 1$ (Moderate), $-1 < b \geq -2$ (Easy), and $b < -2$ (Very Easy) (Dharma, 2024; Hanna & Retnawati, 2022).

The following table presents the recapitulation of Rasch model item difficulty levels from the developed HOTS instrument for fiqh material.

Table 5. Recapitulation of Item Difficulty Levels

Difficulty Criteria	Classification	Item Number
b > 2	Very Difficult	23
1< b ≤ 2	Difficult	10, 13, 17, 31
-1< b ≤ 1	Medium	1, 3, 6, 7, 9, 11, 12, 14, 15, 16, 18, 19, 20, 21, 22, 24, 25, 26, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40
-1< b ≥ -2	Easy	2, 4, 5, 8, 27
b < -2	Very Easy	-

Source: Results of analysis using the Quest program x

Based on Table 5, the 40 developed items show varied difficulty levels: 5 easy items (12.5%), 30 moderate items (75%), 4 difficult items (10%), and 1 very difficult item (2.5%). The 75% proportion of moderate-level items indicates a balanced and student-friendly assessment approach for various student ability levels. This is in line with the HOTS objective of encouraging critical thinking skills, problem solving, and applying knowledge in new situations, because moderate items require more complex thinking processes than easy items (Dewi & Sorensen, 2023).

Analysis of students' ability levels was carried out through the Summary of Case Estimate using the Quest Program, as presented in Table 3. Based on the criteria of Setyawarno (2017) in(Hanna & Retnawati, 2022; Rahayu & Putranta, 2025)Ability categories are divided into three: high (> 1.00), moderate (-1.00 – 1.00), and low (<1.00). The results of the analysis show that the participants' ability was estimated at 0.90, so they fall into the moderate ability category. In addition, this instrument has a high reliability level of 0.90, so the results of the ability assessment are considered consistent and trustworthy.

Item Fit

The determination of whether an item passes (Fit) or does not pass (Misfit) in the Rasch model using the Quest Program can be seen from the OUTFIT t value. Items are included in the fit category if the OUTFIT t value is ≤ 2.00, while items are considered not fit (Misfit) if the OUTFIT t value is ≥ 2.00 (Rahayu & Putranta, 2025). The following table presents the recapitulation of item fit based on the Quest Program.

Table 6. Recapitulation of Item Fit

OUTFIT t Range	Category	Item Number
OUTFIT t ≤ 2,00	Pass (Fit)	2, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, and 40
OUTFIT t ≥ 2,00	Fail (Misfit)	1, 3, 13, and 23

Source: Results of analysis using the Quest program x

Based on the item summary table using the Rasch model, it can be seen that of the 40 developed questions, 36 questions are declared fit, and 4 questions are declared misfit based on the OUTFIT t value. This indicates that most of the developed questions are already in accordance with the Rasch model and are feasible to be used to measure students' higher-order thinking skills. Meanwhile, the questions that are declared not fit (Misfit) need to be revised so that their difficulty levels are appropriate and to ensure clarity and relevance of the material, so that the instrument becomes more accurate and consistent (Muhajir & Hidayat, 2023).

Discussion

The findings of this research, conducted at State Senior High School 9 in Jambi City, indicate that, in terms of quality, the HOTS instrument for Islamic Religious Education, specifically fiqh, demonstrates excellent performance. This conclusion was based on an average Aiken's V coefficient of 0.95. This high figure suggests a significant agreement among the experts in terms of the developed HOTS instrument for fiqh content. Hence, this instrument is considered content-valid and can be used in assessing students' higher-order thinking skills. The results in this study are in accordance with (Mawangir, 2022) who evidenced that the HOTS instrument that has been developed for the fiqh topic proved to have high validity and can be used as an evaluation tool for learning.

In addition to content validity, the quality of the instrument is also seen from its reliability level. Based on the Rasch Model, the item reliability value developed in this study reached 0.96, and the respondent reliability was 0.90. These reliability values indicate that the instrument has very good consistency in measuring students' higher-order thinking skills (Kusumadani et al., 2025). The outcomes of this take a look at are in keeping with the research of Sa'idah & Isnaini, (2020) which states that instruments with high reliability values indicate that the items are consistent in measuring students' abilities, so that the measurement results obtained can be trusted.

The difficulty level of the HOTS instrument for fiqh material developed in this study shows that 75% fall into the moderate category. The majority of items with moderate difficulty levels indicate that the developed instrument has a balanced level of difficulty, which is sufficiently challenging but still reachable by students with average abilities (Mawangir, 2022; Muhajir & Hidayat, 2023). This finding is in line with the research of (Thowiyyah & Hidayatulloh, 2024) which also shows that the difficulty level of items in the developed instrument is mostly in the moderate category, so the instrument is considered effective in measuring students' abilities optimally.

The development of these questions resulted in 36 items that passed (fit), while the other 4 items were declared not to pass (misfit). These results indicate that most items have functioned according to the measurement model used. The presence of several misfit items indicates the need for improvement or revision so that the quality of the instrument becomes more optimal (Yudha, 2023). Instrument revision is carried out to ensure that each item has adequate reliability in measuring students' higher-order thinking skills in fiqh material accurately and comprehensively. Appropriate revisions are expected to improve the consistency and validity of evaluation results, while also ensuring a balanced distribution of difficulty levels in the measurement instrument (Mahmudi et al., 2024; Mawangir, 2022; Muhajir & Hidayat, 2023; Sutiyono, 2022).

Furthermore, the findings of this study also strengthen previous research results that emphasize the need for the development of HOTS-based evaluation instruments in the fiqh subject (Thowiyyah & Hidayatulloh, 2024). The development of the HOTS instrument in this study has advantages compared to several previous studies, because this study uses a fairly large sample and is analyzed using the Rasch Model to evaluate instrument quality. Thus, the measurement of instrument quality is more objective and comprehensive, because it is able to analyze item characteristics and student abilities simultaneously and is relatively sample-independent. The developed instrument not only meets good measurement tool standards, but also has a higher level of measurement accuracy (Mufrihah, 2025) and provides an important methodological contribution to the development of evaluation instruments in Islamic Religious Education, particularly fiqh material. The developed HOTS instrument is expected to provide a real contribution in addressing the limitations of HOTS instruments through the provision of valid and reliable HOTS-based evaluation instruments.

CONCLUSION

This study successfully developed an HOTS instrument in the subject of Islamic Religious Education, fiqh material, and piloted it on twelfth-grade students at State Senior High School 9 in Jambi City. The instrument was of excellent quality. The average Aiken's V coefficient value of 0.95 indicates strong agreement among experts regarding the content of the developed HOTS instrument for fiqh material. The results of the Rasch model analysis using the Quest program show an item reliability value of 0.96 and a person reliability value of 0.90. These reliability values indicate that the instrument has very good consistency in measuring students' higher-order thinking skills. The results of the Rasch model analysis also show that 75% of the developed questions fall into the moderate difficulty category, and there are 36 items declared fit, and 4 items declared not fit. From these findings, it can be concluded that the HOTS instrument in the subject of Islamic Religious Education, fiqh material, that was developed, meets the criteria of being valid, reliable, and feasible to be used as an evaluation tool to measure senior high school students' higher-order thinking skills, although there are still several items that need to be revised. The limitation in the development of this tool lies within the limited scope of fiqh cloth. Therefore, it is expected that future research can develop HOTS fiqh instruments on other materials so that students are trained to think critically, so that students not only understand basic fiqh concepts but are also able to understand and apply fiqh rules in real life.

Acnowledgment

The researcher expresses sincere gratitude to the Education Fund Management Group (LPDP) and the Ministry of Religious Affairs through the MORA THE AIR FUND program for the financial assistance provided for this research. The highest appreciation is also expressed to the Sultan Thaha Saifuddin State Islamic University of Jambi for the guidance and educational facilities provided. Furthermore, the author expresses his deepest gratitude to the schools that participated and assisted in the data collection process for this research.

REFERENCES

Ahmad, N. A., Mayouf, A. A., Elias, N. F., & Mohamed, H. (2024). *Learning management system instrument development based on Aiken's V technique*. *13*(5), 3211–3219. https://doi.org/10.11591/ijere.v13i5.28925

Aldossari, A. T. (2021). Creative Thinking Skills Included in the Content of Evaluation Questions in the Curricula of the Kingdom of Saudi Arabia. *International Journal of Education and Practice*, *9*(3), 520–531. https://doi.org/10.18488/journal.61.2021.93.520.531

Amalia, E. R., M, D. B. I., Khoiriyati, S., & Ummah, N. (2025). *Bridging Educational Reform and Faith : Evaluating Kurikulum Merdeka's Compatibility with Islamic Values in Madrasahs*. *5*, 483–500.

Anggraini, N., Nabillah, C., Lonan, H. D., & Sain, H. (2023). *Application of the Rasch Model to Test Tools In The 3*(1), 19–27.

Ariffin, S. R., Omar, B., Isa, A., & Sharif, S. (2010). *Validity and Reliability of Multiple Intelligent Items Using Rasch Measurement Model*. *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Caronni, A., Ramella, M., Arcuri, P., Salatino, C., Pigini, L., Saruggia, M., Folini, C., Scarano, S., & Converti, R. M. (2023). *The Rasch Analysis Shows Poor Construct Validity and Low Reliability of the Quebec User Evaluation of Satisfaction with Assistive Technology 2 . 0 (QUEST 2 . 0) Questionnaire*.

Dewi, A. N., & Sorensen, B. (2023). *Higher-order Thinking Skills and Metaphor*. *2023*(18), 1–35.

Dharma, S. (2024). *Desain Instrumen Pengukuran Literasi Lingkungan Siswa SMA dengan Menggunakan Pendekatan Model Rasch Instruments Design for Measuring Environmental Literacy of High School Students Using the Rasch Model Approach*. *9*, 218–236. https://doi.org/10.24832/jpnk.v9i2.5126

Firdaus, M. A. (2024). Development of Assessment Instruments for Fiqh Learning Based on Religious Moderation at Madrasah Tsanawiyah in West Java. *Didaktika: Jurnal Kependidikan*, *13*(1), 515–524. https://www.jurnaldidaktika.org/contents/article/view/391

Hanna, W. F., & Retnawati, H. (2022). *Analisis Kualitas Butir Soal Matematika Menggunakan Model Rasch dengan Bantuan Software Quest*. *11*(4), 3695–3704.

Haryani, F. Y., Hidayatullah, M. F., Yusuf, M., & Asrowi. (2021). Problem-Based Learning for Teaching Humanities. *AIP Conference Proceedings*, *020038*(3), 1–9.

Ismail, Parinduri, M. A., & Ibarra, F. P. (2025). Strengthening the Ideology of Islamic Religious Education in the Era of Technological Disruption. *Jurnal Ilmiah Ilmu Terapan Universitas Jambi*, *9*(2), 465–473. https://doi.org/10.22437/jiituj.v9i2.41841

Ju'subaidi, Mujahidin, A., Abdullah, I., & Choirul Rofiq, A. (2025). Students' critical awareness of the internet and social media use as resources for Islamic learning in Indonesian public senior high schools. *British Journal of Religious Education*, *47*(2), 140–155. https://doi.org/10.1080/01416200.2024.2368888

Kamis, A., Hamzah, R., Education, T., Halim, N. A., Keramat, L., & Lumpur, K. (2013). *Competency Assessment of Clothing Fashion Design : Rasch Measurement Model for Construct Validity and Reliability Soaib Asmiran Asimiran Department of Foundation of Education, Faculty of Educational Studies, Universiti*. *5*(2), 1–13.

Karwadi, K., Zakaria, A. R., & Syafii, A. (2024). A Review of the Effects of Active Learning on Higher-Order Thinking Skills: A Meta-Analysis Within Islamic Education. *Journal of Education and Learning*, *18*(1), 97–106. https://doi.org/10.11591/edulearn.v18i1.20895

Kusumadani, A. I., Afandy, H., Agustina, L., Astuti, R., & Waluyo, M. (2025). *Evaluation of Higher-Order Thinking Skills of Middle School Students on Vibration and Wave Topic Using Rasch Measurement*. *11*(5), 74–84. https://doi.org/10.29303/jppipa.v11i5.10900

Kuswanto, A. N. S. & H. (2019). *Developing an Essay Test Instrument for Measuring Diagram Representation and the Capability of Argumentation on Newton's Law*. https://doi.org/10.1088/1742-6596/1227/1/012030

Lestari, W., Wigati, I., Sholeh, M. I., & Pramita, D. (2022). *Instrumen Literasi Digital Guru Menggunakan Model Rasch*. *6*(December), 104–113.

Mahmudi, M. A., Syafruddin, S., Jumahir, J., Haluti, F., Safingah, K., Ilham, I., Syukur, T. A., Inayati, I. N., & Sudirman, S. (2024). Pengantar Pendidikan Agama Islam. In Y. A. Putra (Ed.), *CV Hei Publishing Indonesia*. Hei Publishing Indonesia.

Mawangir, M. (2022). The Development of Higher Order Thinking Skill Test Instrument on the Fiqh Subject: The Case of a State Islamic Senior High School in West Bangka Regency. *Journal of Higher Education Theory and Practice*, *22*(7), 167–177. https://doi.org/10.33423/jhetp.v22i7.5281

Mufrihah, A. (2025). *Rasch Model Analysis of Santri Reverence Morals Scale*. *8*, 1–18.

Muhajir, M., & Hidayat, W. (2023). *Fıqh Subject Exam Questıons Analysıs : Is ıt Based on HOTS*? *44*(5), 868–884.

Rahayu, D. S., & Putranta, H. (2025). *Analisis Kualitas Butir Soal AKM Literasi dan Numerasi Fiqih dengan Model Rasch Berbasis Software Quest*.

Rashed, Z. N., Zhaffar, N. M., Ihwani, S. S., & Iswantir, M. (2025). Driving Islamic Education in the Era of the Industrial Revolution 4.0: Challenges and Strategies. *Global Journal Al-Thaqafah*, *15*(1), 18–33. https://doi.org/10.7187/GJAT072025-2

Ridwan, M. R., Hadi, S., & Jailani, J. (2023). *Measurement of Psychometric Properties: Numerical Aptitude Assessment Scale for Prospective High School Students : A Rasch Model Analysis*. *12*(4), 2416–2429. https://doi.org/10.18421/TEM124

Sa'idah, N., & Isnaini, I. K. (2020). Pengembangan Instrumen Tes Higher Order Thinking Skill (HOTS) Mata Pelajaran Fikih Kelas XI di MA Masalikil Huda Tahunan Jepara. *Al-Tadzkiyyah: Jurnal Pendidikan Islam*, *11*(2), 209–220.

Sihombing, B. (2024). *Model Pengembangan 4D (Define, Design , Develop , dan Disseminate) dalam Pembelajaran Pendidikan Islam*. *4*, 11–19.

Sugiyono. (2021). *Metode Penelitian Pendidikan: Kuantitatif, Kualitatif, Kombinasi, R&D dan Penelitian Pendidikan*. Alfabeta.

Sutiyono, A. (2022). Developing the Instrument of Fikih Assessment in Madrasah Aliyah. *Nadwa: Jurnal Pendidikan Islam*, *16*(1), 104–122. https://doi.org/10.21580/nw.2022.16.1.15384

Syahidan, J., & Mukminin, A. (2024). *Analisis Kesulitan Belajar Siswa dalam Memahami Materi Fiqih pada Kelas 9 MTsN 2 Wonogiri Analysis of Students ' Learning Difficulties in Understanding Fiqh Materials in Grade*. *6*(2), 1–13.

Tatang Hidayat, S. (2019). Inovasi Pembelajaran Pendidikan Agama Islam Melalui Model Contextual Teaching and Learning. *Pendidikan Agama Islam*, *16*(2), 115–136. https://ejournal.uin-suka.ac.id/tarbiyah/jpai/article/view/2631

Thowiyyah, H., & Hidayatulloh, M. K. Y. (2024). Development of a HOTS-based Multiple Choice Question Test Instrument on the Basic Competency of Understanding Marriage Provisions According to Legislation. *Application: Applied Science in Learning Research*, *3*(3), 13–21. https://doi.org/10.32764/application.v3i3.4549

Tolchah, M., & Mu'ammar, M. A. (2019). *Islamic Education in The Globalization Era : Challenges, Opportunities, And Contribution of Islamic Education In Indonesia*. *7*(4), 1031–1037.

Tuli, T., & Munirah. (2022). Efektivitas Strategi Problem-Based Learning (PBL) terhadap Hasil Belajar Fiqih di Madrasah Aliyah Al-Falah Limboto Barat. *PEKERTI: Journal Pendidikan Islam & Budi Pekerti*, *4*(1), 19–29. https://journal.iaingorontalo.ac.id/index.php/pekerti/article/view/2901

Wahyuni, V. (2023). *Aiken Index Analysis to Determine the Validity of the Contents of the Project Assessment Instrument in Learning Islamic Religion and Character in the Independent Curriculum*. *6*(2), 565–569.

Wiliyanti, V., Rakhmadi, R., & Octavia, S. (2023). Islamic Religious Education on the Technological Developments of the Era-Revolutionary Society 5.0. *KnE Social Sciences*, *2023*, 647–657. https://doi.org/10.18502/kss.v8i16.14076

Yang, S., Tsou, M., Chen, E., & Chan, K. (2011). Statistical Item Analysis of the Examination in

Anesthesiology for Medical Students using the Rasch Model. *Journal of the Chinese Medical Association*, *74*(3), 125–129. https://doi.org/10.1016/j.jcma.2011.01.027

Yıldırım, H., & Gülengül, E. (2025). Grounding the Idea of Renewal in Islamic Thought, and Collective Iḥyā': A Study from the Perspectives of the Philosophy of Religion and Fiqh. *Humanities and Social Sciences Communications*, *12*(1), 1–10. https://doi.org/10.1057/s41599-025-05198-9

Yudha, R. P. (2023). *Higher Order Thinking Skills (HOTS) Test Instrument : Validity and Reliability Analysis with The Rasch Model*. *12*(1), 21–38.

Zeileis, A. (2025). *Examining Exams Using Rasch Models and Assessment of Measurement Invariance*. *54*, 9–26. https://doi.org/10.17713/ajs.v54i3.2055