

Evaluation of Islamic Religious Education Learning on SKI Material: Development of HOTS Instrument in Senior High School

Bayu Lesmana¹, Risnita^{2*}, Dedi Sastradika², Sri Ramdayeni Sakunti²

¹Islamic Religious Education, UIN Sultan Thaha Saifudin Jambi, Indonesia

²UIN Sultan Thaha Saifudin Jambi, Indonesia

✉ Author Corresponding: risnita@uinjambi.ac.id*

ABSTRACT

Learning of Islamic Religious Education (PAI) on the subject of Islamic Cultural History (SKI) is still dominated by conventional methods and evaluation that has not been oriented toward Higher Order Thinking Skills (HOTS), resulting in students' critical thinking skills not developing optimally. This study aims to develop and validate a Higher Order Thinking Skills (HOTS)-based assessment instrument for the Islamic Cultural History (SKI) subject at the eleventh-grade level of senior high school. The research employed a Research and Development (R&D) approach using a modified 4D model consisting of define, design, and develop stages. The instrument comprised 30 multiple-choice items measuring cognitive levels C4, C5, and C6 based on the revised Bloom's taxonomy. Content validity was analyzed using Aiken's V, while reliability, item difficulty, and discrimination were examined using the Rasch model through QUEST program software. The results showed a very high content validity index (Aiken's V = 0.88) and excellent reliability (0.91). Rasch analysis indicated that most items fit the model (INFIT MNSQ 0.77–1.33), with a balanced distribution of item difficulty and very good discrimination indices. Three items were identified as misfit and subsequently revised to maintain the composition of 30 HOTS-based items. This study contributes to the development of a psychometrically tested HOTS assessment instrument for Islamic Religious Education, integrating Rasch modeling to ensure objectivity and measurement precision. The instrument provides a more robust and theoretically grounded evaluation tool compared to previous studies that relied on classical test analysis or did not fully integrate HOTS constructs.

Keywords: HOTS Assessment; Islamic Cultural History; Rasch Model; Instrument Development; PAI Evaluation.



Article History:

Received: 08-02-2026

Revised : 02-07-2026

Accepted: 03-07-2026

Online : 01-08-2026

How to Cite (APA style):

Lesmana, B., Risnita, Sastradika, D., & Sakunti, S. R. (2026). Evaluation of Islamic Religious Education Learning on SKI Material: Development of HOTS Instrument in Senior High School. *IJECA (International Journal of Education and Curriculum Application)*, 9(2), 590-600. <https://doi.org/10.31764/ijece.v9i2.38225>



This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license

1. INTRODUCTION

The role of Islamic Religious Education (PAI) learning in the era of information complexity and various global challenges is not limited to the transfer of knowledge, but also includes the cultivation of critical thinking and strong moral reasoning. However, Islamic Religious Education still faces significant challenges related to teaching methods that tend to be conventional and teacher-centered (Hussien et al., 2021). Various studies show that classroom learning processes are still dominated by traditional approaches, which make the learning atmosphere monotonous and less engaging for students (Tabroni et al., 2022). This problem is exacerbated by various

factors, ranging from teachers' limited digital competence (Reksiana et al., 2024), the lack of innovation in media use (Mazidah et al., 2023), to inadequate resources and supporting facilities (Badriyah et al., 2025). As a result, PAI subjects, especially SKI material, are often perceived as difficult and merely focused on memorization (Assalihee et al., 2024).

The challenge of conventional teaching is particularly evident in the subject of Islamic Cultural History (SKI). Specifically, the SKI material has characteristics that require contextual presentation to be relevant for students (Arif & Aziz, 2023). The narrative nature of SKI material often faces obstacles in connecting the values and lessons (*ibrah*) from historical events with the realities of modern life faced by students (Tabroni et al., 2022). This challenge becomes even greater due to the limited time allocation for learning, making it difficult for teachers to deliver the material comprehensively and in depth (Suntiah, 2021). When historical content is presented merely as a series of facts without efforts at contextualization, students become less interested and fail to grasp the essential lessons from Islamic civilization (Arif & Aziz, 2023).

The combination of monotonous teaching methods and the lack of contextual presentation directly impacts the low development of students' Higher-Order Thinking Skills (HOTS) (Karwadi et al., 2024). A learning approach focused on memorization hinders students from developing analytical, evaluative, and creative abilities (Assalihee et al., 2024). Various studies emphasize that the implementation of HOTS in PAI and SKI learning remains very limited and has not yet been optimal (Hamdi et al., 2022; Jima'ain et al., 2024). Often, teachers' understanding of HOTS is limited to constructing question models, rather than viewing it as an integrated approach within the entire teaching-learning process (Kosasih et al., 2022).

As a result of these various problems, students are not trained to think critically, solve problems, or reflect on Islamic values from the materials they study (Hussien et al., 2021; Suntiah, 2021). Learning becomes less meaningful because students are only able to recall information without being able to process and apply it in daily life (Mazidah et al., 2023). The low level of HOTS ability ultimately affects not only cognitive learning outcomes but also fails to develop character and critical reasoning skills that align with the main objectives of Islamic education, in which assessment should not only focus on the cognitive aspect alone (Nasucha et al., 2023).

Although previous studies have developed assessment instruments in Islamic Religious Education, several limitations remain. Some studies produced valid and reliable instruments but did not integrate HOTS-oriented constructs in the item design (Sutiyono, 2022). Other studies developed HOTS-based instruments but involved relatively small sample sizes or relied on classical test analysis, which limits the precision and objectivity of measurement (Dewi et al., 2023; Sa'idah & Isnaini, 2020). In addition, previous studies have not comprehensively combined large-scale trials with Rasch model analysis to ensure the psychometric quality of HOTS instruments in SKI learning.

Based on this synthesis, the research gap lies in the absence of a psychometrically robust HOTS instrument for SKI that is developed through a systematic model, tested on a large sample, and analyzed using the Rasch measurement approach. Therefore, this study aims to develop and validate a HOTS-based assessment instrument for SKI by examining: (1) content validity using expert judgment; (2) reliability using the Rasch model; (3) item difficulty distribution; and (4) item discrimination indices. This study is expected to contribute theoretically to the development of HOTS-based assessment in Islamic Religious Education and methodologically by demonstrating the application of the Rasch model for instrument validation in the SKI context.

2. RESEARCH METHOD

This study employed a Research and Development (R&D) approach using a modified 4D model consisting of three stages: Define, Design, and Develop. The Dissemination stage was not implemented due to time and resource limitations; therefore, the model was adapted into a 3D procedure. This modification does not affect the core development process because the main objective of this study was limited to instrument construction and psychometric validation. The modified 4D development model is presented in Figure 1.

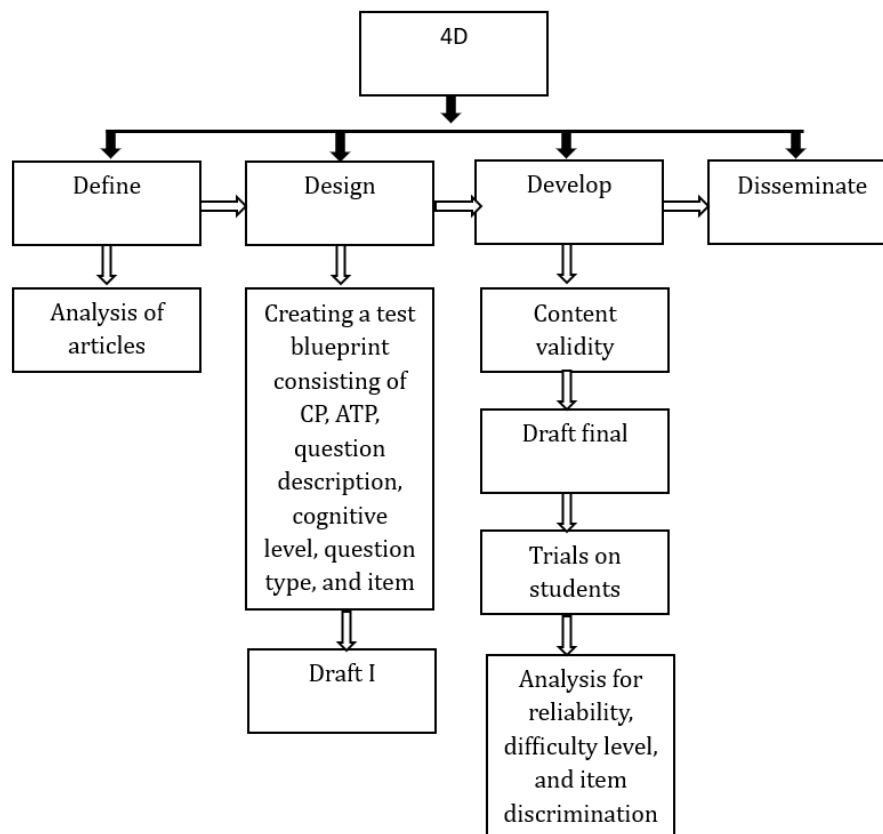


Figure 1. 4D model

The Define stage aims to determine the product to be developed along with its specifications. This stage involves a needs analysis conducted through research and literature review. The second stage, Design, involves creating a design for the predetermined product. The next stage, Develop, includes activities to transform the design into a product and test its validity repeatedly until a product that meets the specified standards is produced. The final stage, Dissemination, involves disseminating the tested product so that it can be utilized by others. However, the Dissemination stage was not carried out in this study due to time and resource limitations.

The research subjects involved one lecturer of Islamic Religious Education (PAI) at UIN STS Jambi as a validator to assess material suitability; one lecturer in learning evaluation at UIN STS Jambi as a validator to assess the suitability of the items with HOTS indicators; one Indonesian language lecturer at UIN STS Jambi as a validator to assess language rule appropriateness; one PAI lecturer to assess the accuracy of Qur'anic verses; as well as a PAI subject teacher and peer reviewers. Furthermore, the questions were tested with 300 eleventh-grade students from several senior high schools in Jambi Province. The sampling technique used was cluster random

sampling, in which several schools were selected to represent different academic backgrounds. The respondents consisted of students aged 16–17 years who had studied the SKI material in accordance with the Merdeka Curriculum.

The assessment instrument developed was a multiple-choice test for the subject of Islamic Religious Education and Character Education, specifically focusing on Islamic Cultural History material. The instrument consisted of 30 items. The questions were designed based on the revised Bloom's Taxonomy indicators, including 15 items at level C4 (analyzing), 10 items at level C5 (evaluating), and 5 items at level C6 (creating). This instrument is expected to measure students' higher-order cognitive abilities in analyzing, evaluating, and creating within PAI learning.

The data analysis techniques used were both qualitative and quantitative. Qualitative data analysis was obtained from literature studies to identify needs. Quantitative data analysis included content validity testing using the Aiken index. Validity testing was assessed based on expert validation results. If the score is < 0.4 , it is considered less valid; a score of $0.4\text{--}0.8$ indicates moderate validity; and it is considered very valid if the index is > 0.8 (Aiken, 1985). Qualitative data analysis also uses the Rasch measurement model with QUEST software. The main indicators examined included item fit statistics (INFIT MNSQ), reliability coefficients, item difficulty (logit values), and point-biserial discrimination indices. Items were considered fit if the INFIT MNSQ value ranged between 0.77 and 1.33.

The Rasch model is an important measurement tool because it formalizes the invariant requirement for data, ensuring that the instrument functions consistently across different sample groups, thereby increasing the validity of comparisons (Hagquist & Andrich, 2017). This model allows independent estimation of item and person parameters, placing them on a common logit scale, thus making it more meaningful than traditional assessment methods (Hagquist & Andrich, 2017; Tesio et al., 2024).

3. RESULTS AND DISCUSSION

3.1 Define Stage

This stage began with a document review in the form of an analysis of articles discussing problems in Islamic Religious Education (PAI) learning, particularly in PAI learning evaluation. From the results of the document analysis, it was found that PAI learning evaluation had not yet used HOTS instruments, resulting in low students' critical thinking skills. The next step was the document review, which included the Learning Outcomes (CP) and Learning Objective Flow (ATP) referring to the Merdeka Curriculum. This was conducted to create an instrument that is relevant to and based on the applicable curriculum.

3.2 Design Stage

The design stage began by creating a test blueprint consisting of CP, ATP, question description, cognitive level, question type, and item formulation. The questions developed referred to the revised Bloom's Taxonomy. According to the revised Bloom's Taxonomy, HOTS questions start from level C4 (analyzing), C5 (evaluating), and C6 (creating). The instrument developed must be based on the framework of the revised Bloom's Taxonomy and must also align with the Learning Outcomes (CP) and Learning Objective Flow (ATP). This ensures that the instrument truly measures students' higher-order thinking abilities.

3.3 Develop Stage

The designed instrument was then tested in the development stage. This stage consisted of two steps: content validity testing conducted by experts and trials on students. The validity test was carried out by three lecturers from UIN STS Jambi who are experts in their respective fields, as well as a PAI subject teacher with over ten years of experience and peer reviewers. The aspects evaluated in the validity test included language, accuracy of verses/hadith, relevance of issues, and alignment of questions with the HOTS level. The validation sheet used a four-point scale: a score of 1 (poor), 2 (fair), 3 (good), and 4 (very good). The instrument was then tested on 300 students, and the results were analyzed for reliability, difficulty level, and item discrimination.

a. Validation Results

The validation results obtained from five validators including three lecturers from UIN STS Jambi, one Islamic Religious Education teacher, and one peer reviewer are presented in Table 1 below:

Table 1. Content Validation Results

No. Item	Average	Aiken V	Category	No. Item	Average	Aiken V	Category
b1.	3.6	0.87	Very Valid	b16	3.6	0.87	Very Valid
b2.	3.6	0.87	Very Valid	b17	3.8	0.93	Very Valid
b3.	3.8	0.93	Very Valid	b18	3.6	0.87	Very Valid
b4.	3.8	0.93	Very Valid	b19	3.8	0.93	Very Valid
b5.	3.8	0.93	Very Valid	b20	3.6	0.87	Very Valid
b6.	4.0	1.00	Very Valid	b21	3.6	0.87	Very Valid
b7.	3.8	0.93	Very Valid	b22	3.6	0.87	Very Valid
b8.	4.0	1.00	Very Valid	b23	3.6	0.87	Very Valid
b9.	3.8	0.93	Very Valid	b24	3.6	0.87	Very Valid
b10.	3.6	0.87	Very Valid	b25	3.4	0.80	Very Valid
b11.	3.6	0.87	Very Valid	b26	3.6	0.87	Very Valid
b12.	3.6	0.87	Very Valid	b27	3.4	0.80	Very Valid
b13.	3.6	0.87	Very Valid	b28	3.4	0.80	Very Valid
b14.	3.6	0.87	Very Valid	b29	3.4	0.80	Very Valid
b15.	3.8	0.93	Very Valid	b30	3.4	0.80	Very Valid
Aiken's V Average						0,88	Very Valid

The HOTS-based test instrument for Islamic Religious Education, particularly in the Islamic Cultural History material for Grade XI Senior High School, underwent content validation. The validation by five experts included the aspects of language, alignment with verses/hadith, relevance of issues, and question alignment with the HOTS level. The validation results showed an average Aiken's V value of 0.88. This value falls within the "very valid" category (Aiken, 1985). Although the validation score indicated very good results, the suggestions and feedback from the experts were important for perfecting the instrument. For example, several questions needed to be adjusted to better match the HOTS level.

b. Item Fit Analysis

The analysis of item fit using the Rasch model can be seen from the INFIT MNSQ value. The overall item fit was generally good, with an average INFIT MNSQ value of 1.00. However, there were items classified as misfit because they fell outside the ideal range of 0.77–1.33. Underfit (MNSQ < 0.77) indicates that the item is too easy or has low discrimination

(Makmur et al., 2025). The trial data analysis showed that most items were within the ideal range, although a few were outside it. The range of INFIT MNSQ scores is explained in Table 2.

Table 2. INFIT MNSQ Values

Item	INFIT MNSQ	Category	Item	INFIT MNSQ	Category	Item	INFIT MNSQ	Category
b1.	1.24	Pass	b11	0.95	Pass	b21	1.25	Pass
b2.	0.73	Fail	b12	1.16	Pass	b22	0.74	Fail
b3.	0.95	Pass	b13	1.15	Pass	b23	1.05	Pass
b4.	0.81	Pass	b14	0.97	Pass	b24	0.94	Pass
b5.	0.83	Pass	b15	0.86	Pass	b25	1.11	Pass
b6.	0.85	Pass	b16	0.84	Pass	b26	1.02	Pass
b7.	0.80	Pass	b17	0.79	Pass	b27	1.29	Pass
b8.	1.43	Fail	b18	0.92	Pass	b28	0.85	Pass
b9.	0.95	Pass	b19	0.97	Pass	b29	1.06	Pass
b10.	1.02	Pass	b20	1.15	Pass	b30	1.13	Pass

From the table above, it can be concluded that the majority of the developed items fall within the acceptable range. Three items were eliminated; for example, item 2 (0.73) and item 22 (0.74) were underfit, meaning they were too easy or had low discrimination. Overfit (MNSQ > 1.33) indicates that the response pattern deviated from the Rasch model, suggesting excessive guessing or misconceptions, for example, item 8 (1.43). It can be concluded that the development of the HOTS-based test instrument in Islamic Religious Education, particularly in Islamic Cultural History for Grade XI Senior High School, demonstrated good item discrimination, with only three items requiring revision. Three items identified as misfit were revised based on expert feedback and Rasch output analysis. The revisions focused on improving distractor effectiveness and aligning the cognitive demand with HOTS indicators. After revision, the total composition of the instrument remained 30 items, ensuring proportional representation across cognitive levels C4, C5, and C6.

c. Reliability Analysis

The validated items were then tested to assess reliability. Reliability can be defined as trustworthiness, consistency, dependability, or stability (Zijlmans et al., 2017). Good reliability is indicated by the consistency and stability of results obtained from an instrument when used under similar conditions at different times. A reliable instrument will yield the same or very similar results when used to measure the same construct under comparable conditions (Kimberlin & Winterstein, 2008). The HOTS-based test instrument for Islamic Religious Education, particularly for the Islamic Cultural History material for Grade XI Senior High School, obtained a reliability score of 0.91. The reliability value is presented in Table 3.

Table 3. Reliability Score

Summari of item Estimate	Value
Mean	0.00
SD	0.68
SD (adjusted)	0.65
Reliability of estimate	0.91

This value is categorized as highly reliable according to the Rasch model (Boone et al., 2014). This proves that the developed instrument possesses consistency and dependability as a measurement tool, indicating that it can produce stable and consistent results if administered under similar conditions.

d. Difficulty Level Values

The item difficulty index is categorized as Very Difficult if the value ≥ 0.53 , Difficult if between $0.52 - 0.0$, Easy if between $0.0 - (-0.52)$, and Very Easy if ≤ -0.54 (Makmur et al., 2025). Based on the analysis results, the HOTS-based test instrument in Islamic Religious Education, particularly in the Islamic Cultural History material for Grade XI Senior High School, obtained an overall score of 0.68, which is categorized as difficult. The item difficulty values are presented in Table 4 below:

Table 4. Difficulty Level Values

Item	Difficulty	Category	Item	Difficulty	Category
b1	-1.13	Very Easy		0.17	Difficult
b2	-1.13	Very Easy		-0.37	Easy
b3	-0.37	Easy		-0.24	Easy
b4	-0.76	Very Easy		0.69	Very Difficult
b5	-0.05	Easy		0.61	Difficult
b6	-0.08	Easy		1.39	Very Difficult
b7	-0.88	Very Easy		1.15	Very Difficult
b8	1.03	Very Difficult		-0.61	Very Easy
b9	-0.05	Easy		-0.56	Very Easy
b10	-0.24	Easy		-0.56	Very Easy
b11	0.03	Difficult		-0.37	Easy
b12	0.55	Difficult		0.20	Difficult
b13	0.96	Very Difficult		1.03	Very Difficult
b14	-0.71	Very Easy		0.10	Difficult
b15	-0.08	Easy		0.26	Difficult

In detail, the developed test items show a wide distribution of difficulty levels, ranging from very easy items such as items 1, 2, 4, 7, 14, 23, 24, and 25; easy items such as items 3, 5, 6, 9, 10, 15, 17, 18, and 26; difficult items such as items 11, 12, 13, 16, 27, 29, and 30; and very difficult items such as items 8, 19, 20, 21, 22, and 28. These results illustrate that the developed instrument maintains a balanced distribution among easy, moderate, and difficult questions. A good test instrument for evaluating learning outcomes generally has a wide and balanced spread of question difficulty levels, ranging from very easy to very difficult.

e. Discrimination Power Analysis

The discrimination power analysis in the Rasch model can be seen from the point biserial value. According to Reynolds et al. (2009), the discrimination power has score ranges as follows: ≥ 0.40 (very good), $0.30-0.39$ (good), $0.11-0.29$ (moderate), and $0.00-0.10$ (poor). Based on the results of the discrimination power analysis, the obtained scores are presented in Table 5.

Table 5. Discrimination Power

Score Range	Item Number	Category
≥ 0.40	2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 22, 23, 24, 25, 26, 28, 29, 30	Very Good
0.30 - 0.39	1, 8, 21, 27	Good
0,11 - 0,29	-	-
0.00 - 0.10	-	-

Based on the data analysis results, the quality of the test items shows a very good discrimination power. A total of 86.6% of the items are in the Very Good category, while the remaining 13.4% are in the Good category. There are no items categorized as Moderate or Poor, indicating that overall, the items are highly representative for measuring students' abilities.

3.4 Disseminate Stage

The disseminate stage is the final phase in the 4D development model. In this stage, the instrument that has been tested and refined should be socialized, printed, and distributed to educational institutions or organizations (Alvionita et al., 2020). However, in this study, the researcher only conducted the research up to the third stage, namely Develop, due to time and resource limitations.

3.5 Discussion

This study successfully developed and validated 30 Higher Order Thinking Skills (HOTS)-based test items for the Islamic Cultural History (SKI) material in Grade XI Senior High School. The results indicated positive findings, marked by a content validity value of 0.88 (very valid category) and a reliability value of 0.91. These findings demonstrate that the developed instrument aligns with the Learning Outcomes (CP) and Learning Objective Flow (ATP), and is capable of measuring students' higher-order thinking skills at cognitive levels C4, C5, and C6. The discrimination power and difficulty level of the test items also showed good categories, although three items required revision or elimination based on the Rasch model analysis.

The high validity and reliability of the instrument can be attributed to the systematic development process using the 4D model, particularly in the Define and Design stages, which emphasized needs analysis, curriculum alignment, and formulation of HOTS indicators based on the revised Bloom's Taxonomy. The involvement of expert validators from various fields also ensured the appropriateness of content, language, and item construction.

Furthermore, the difficulty level analysis results showed a wide distribution of items, ranging from very easy to very difficult. Items categorized as very easy occurred because the distractor options were ineffective, as indicated when several distractor options were chosen by fewer than 5% of total students. Conversely, items categorized as very difficult occurred when the incorrect options functioned effectively as distractors, as shown when some distractors were selected by more than 5% of total students. However, as emphasized by Pratama et al. (2021), a good evaluation instrument should have a balanced distribution of difficulty levels. The total of 300 respondents also served as a supporting factor for producing a high-quality evaluation instrument. This aligns with the findings of Shimada & Katahira (2023), who stated that the number of respondents affects trial results.

The results of this study are consistent with the findings of Dewi et al. (2023), which reported that HOTS-based assessment instruments in PAI and SKI learning demonstrated high validity and reliability, although those studies involved fewer respondents. However, this study presents a significant difference compared to previous research, particularly that of Sutyono (2022), which did not integrate HOTS-based questions, and Sa'idah & Isnaini (2020) which still used classical test models for data analysis. By integrating the Rasch model analysis, this research offers an advantage over previous studies because the Rasch model allows for a more comprehensive analysis of item quality, including reliability, difficulty level, and item discrimination power.

Differences in educational contexts, development models, item analysis techniques, and respondent numbers make this study yield new findings that not only confirm the feasibility of the HOTS instrument but also highlight the superiority of the Rasch model in ensuring the quality, stability, and objectivity of measuring students' higher-order thinking abilities in the Islamic Cultural History material. By providing an assessment instrument that is more objective and oriented toward the development of students' critical thinking, this instrument can serve as a reference for teachers and curriculum developers in designing evaluations that align with 21st-century learning demands and the goals of Islamic education. Although this study produced positive results, future research is recommended to proceed to the Disseminate stage so that the instrument can be tested and implemented more broadly across various schools. In addition, further research can develop HOTS instruments based on Computer-Based Testing (CBT) or examine the effectiveness of using this instrument in improving students' critical thinking skills.

4. CONCLUSION

This study successfully developed a HOTS-based assessment instrument for Islamic Cultural History learning at the senior high school level through a systematic and psychometrically grounded process. The instrument demonstrated strong content validity, high reliability, and appropriate item characteristics based on Rasch model analysis, indicating that it is suitable for measuring students' higher-order thinking skills. Theoretically, this study contributes to the development of assessment models in Islamic Religious Education by integrating HOTS constructs with modern psychometric approaches, particularly the Rasch model. Methodologically, it demonstrates the importance of combining systematic development procedures with large-scale trials and item response analysis to ensure instrument quality. Practically, the developed instrument can serve as a reference for teachers and curriculum developers in designing more objective, valid, and critical-thinking-oriented evaluations in SKI learning. Future studies are recommended to proceed to the dissemination stage and test the effectiveness of the instrument across broader educational contexts.

ACKNOWLEDGMENT

The researcher would like to express the deepest gratitude to the Education Fund Management Institution (LPDP), in collaboration with the Ministry of Religious Affairs through the MoRA The Air Fund research program, for the financial support that greatly contributed to this study. The highest appreciation is also extended to the State Islamic University Sulthan Thaha Saifuddin Jambi for its academic support and facilities provided. Furthermore, the author expresses sincere thanks to the participating schools for their cooperation and assistance in the data collection process. Lastly, heartfelt gratitude is given to everyone involved in this study, whose contributions ensured the smooth completion of this research.

REFERENCES

- Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45(1), 131–142. <https://doi.org/doi.org/10.1177/0013164485451012>
- Alvionita, D., Subchan, W., & Iqbal, M. (2020). The development of ecosystem education game based on Baluran National Park for senior high The development of ecosystem education game based on Baluran National Park for senior high school. *Journal of Physics: Conference Series*, 1465(1). <https://doi.org/10.1088/1742-6596/1465/1/012039>
- Arif, M., & Aziz, M. K. N. A. (2023). Islamic Religious Education Learning Model in the 21. *Indonesian Journal of Islamic Education Studies (IJIES)*, 6(2), 237–262. <https://doi.org/10.33367/ijies.v6i2.4417>
- Assalihee, M., Bakoh, N., Boonsuk, Y., & Songmuang, J. (2024). Transforming Islamic Education through Lesson Study (LS): A Classroom-Based Approach to Professional Development in Southern Thailand. *Education Sciences*, 14(9). <https://doi.org/10.3390/educsci14091029>
- Badriyah, L., Yorman, & Wardi, M. (2025). Policy analysis of Islamic educational institutions in facing the challenges of society 5.0: Innovation, learning, and technology-based infrastructure. *Multidisciplinary Reviews*, 9(1), 1–10. <https://doi.org/10.31893/multirev.2026012>
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer. <https://doi.org/10.1007/978-94-007-6857-4>
- Dewi, Y., In'ami, M., Audia', A., & Masrukhin. (2023). Assessment Based on Higher Order Thinking Skill on SKI Learning. *AL-ISHLAH: Jurnal Pendidikan*, 15(1), 443–458. <https://doi.org/10.35445/alishlah.v15i1.2586>
- Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of differential item functioning in health research using the Rasch model. *Health and Quality of Life Outcomes*, 1–8. <https://doi.org/10.1186/s12955-017-0755-0>
- Hamdi, S., Wahdah, N., Ahmadi, & Musthofa, K. (2022). The Problem Of Implementation Of Islamic Education Curriculum In The Aspect Of Moderation Learning And Hots Evaluation. *EDURELIGIA: Jurnal Pendidikan Agama Islam*, 06(01), 21–36. <https://doi.org/10.33650/edureligia.v6i1.3620>
- Hussien, S., Wahab, M. K. A., & Hashim, R. (2021). Improving Students' Inquiry Skills in Islamic Education Through Hikmah Pedagogy and Community of Inquiry. *Malaysian Journal of Learning and Instruction*, 18(2), 189–214. <https://doi.org/10.32890/mjli2021.18.2.7>
- Jima'ain, M. T. A., Mohamed, A. M., Hehsan, A., Saidalvi, A., Nasi, B. M., Faisal, M. S., & Awae, F. (2024). Embedding higher order thinking skills in Islamic history (Sirah) education in Malaysia. *International Journal of Evaluation and Research in Education*, 13(2), 952–959. <https://doi.org/10.11591/ijere.v13i2.26431>
- Karwadi, K., Zakaria, A. R., & Syafii, A. (2024). A review of the effects of active learning on high 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>
- Kimberlin, C. I., & Winterstein, almut G. (2008). Validity and reliability of measurement instruments used in research. *American Society of Health-Systems Pharmacy*, 65(23), 2276–2284. <https://doi.org/10.2146/ajhp070364>
- Kosasih, A., Supriyadi, T., Firmansyah, M. I., & Rahminawati, N. (2022). Higher-Order Thinking Skills in Primary School: Teachers' Perceptions of Islamic Education. *Journal of Ethnic and Cultural Studies*, 9(1), 56–76. <https://doi.org/10.29333/ejecs/994>
- Makmur, F., Retnawati, H., & Arliani, E. (2025). Development and evaluation of cognitive test instruments of sequence and series topic grade XI: The Rasch Model. *AIP Conference Proceedings*, 3285(1). <https://doi.org/10.1063/5.0262927>
- Mazidah, R., Romadlon, D. A., & Astutik, A. P. (2023). The Application of the Snowball Throwing Learning Model Based on HOTS Questions in Improving Islamic Religious Education Learning Outcomes at Junior High School. *Scaffolding: Jurnal Pendidikan Islam Dan*

- Multikulturalisme*, 5(1), 234–250. <https://doi.org/10.37680/scaffolding.v5i1.2086>
- Nasucha, M. R., Khozin, K., & Thoifah, I. (2023). Synergizing Islamic Religious Education and Scientific Learning in the 21st Century: A Systematic Review of Literature. *Jurnal Pendidikan Agama Islam (Journal of Islamic Education Studies)*, 11(1), 109–130. <https://doi.org/10.15642/jpai.2023.11.1.109-130>
- Pratama, D. N., Pratiwi, O. N., & Sutoyo, E. (2021). Classification of Questions Based on Difficulty Levels using Support Vector Machine and Naïve Bayes Algorithms for Imbalanced Class. *Institute of Electrical and Electronics Engineers Inc*, 40–45. <https://doi.org/10.1109/IC2IE53219.2021.9649149>
- Reksiana, Nata, A., Rosyada, D., Rahiem, M., Dinia, H., Rafikjon, U., & Abdulbosit, R. (2024). Digital Extension of Digital Literacy Competence for Islamic Religious Education Teachers in the Era of Digital Learning. *Jurnal Pendidikan Agama Islam*, 21(2), 402–420. <https://doi.org/10.14421/jpai.v21i2.9719>
- Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and Assessment in Education* (2nd ed.). Pearson Education.
- Sa'idah, N., & Isnaini, I. K. (2020). Development of Higher Order Thinking Skill (HOTS) Test Instruments for Fiqh Subjects in Grade XI at MA Masalikil Huda Tahunan Jepara. *Al-Tadzkiyyah: Jurnal Pendidikan Islam*, 11(2), 209–220. <https://doi.org/10.24042/atjpi.v11i2.6889>
- Shimada, D., & Katahira, K. (2023). Sequential dependencies of responses in a questionnaire survey and their effects on the reliability and validity of measurement. *Behavior Research Methods*, 55Pratama,(6), 3241–3259. <https://doi.org/10.3758/s13428-022-01943-z>
- Suntiah, R. (2021). Students' Critical Thinking Skills in the Reflective Class of Islamic Cultural History. *Jurnal Pendidikan Islam*, 7(2), 195–204. <https://doi.org/10.15575/jpi.v7i2.15060>
- 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>
- Tabroni, I., Susana, S., Mulyadi, A., & Zaelani, N. (2022). Utilization of the Discovery Learning Model to Overcome Islamic Cultural History Learning Problems in Madrasa. *Indonesian Journal of Islamic Education Studies (IJIES)*, 5(1), 81–94. <https://doi.org/10.33367/ijies.v5i1.2409>
- Tesio, L., Caronni, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 1. The “most likely” measures coming from the model. *Disability and Rehabilitation*, 46(3), 591–603. <https://doi.org/10.1080/09638288.2023.2169771>
- Zijlmans, E. A. O., Tijmstra, J., & Sijtsma, K. (2017). Item-Score Reliability in Empirical-Data Sets and Its Relationship With Other Item Indices. *Educational and Psychological Measurement*, 1–23. <https://doi.org/10.1177/0013164417728358>