

Developing a Higher-Order Thinking Skills Assessment Instrument for *Akidah Akhlak* Learning: An ADDIE and Rasch Model Approach

Thoiroh¹, Risnita², Dedi Sastradika³, Sri Ramdayeni Sakunti⁴

^{1,2,3,4} Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi, Indonesia; thoi0108.2004@gmail.com

ARTICLE INFO

Keywords:

higher-order-thinking skills;
assessment instrument;
akidah akhlak,
Rasch model;
Islamic education

Article history:

Received 2026-03-05
Revised 2026-08-26
Accepted 2026-09-30

ABSTRACT

The assessment of Higher-Order Thinking Skills (HOTS) in *Akidah Akhlak* learning remains dominated by lower-order cognitive questions, limiting the evaluation of students' analytical, evaluative, and creative thinking abilities. This study aimed to develop and evaluate the psychometric quality of a HOTS-based assessment instrument for Madrasah Aliyah students. This research employed a Research and Development approach using the ADDIE model, integrated with Rasch analysis. The instrument comprised 40 multiple-choice items targeting cognitive levels C4, C5, and C6. Content validity was assessed by five experts using Aiken's V. Following a preliminary trial involving 30 students, field testing was conducted with 300 students from five madrasahs in Jambi Province. Item fit, reliability, and difficulty were analyzed using the QUEST program. Expert validation yielded an average Aiken's V of 0.93, indicating strong content validity. Rasch analysis identified one misfitting item, resulting in a final instrument of 39 items. All retained items satisfied the reported Outfit t criterion (≤ 2.00). Item reliability reached 0.97, while person reliability was 0.69. The mean student ability (2.05 logits) exceeded the mean item difficulty (0.00 logits), indicating that the instrument was relatively easy for the sampled students. The instrument demonstrates promising content validity and item-level psychometric properties for assessing HOTS in *Akidah Akhlak* learning. However, moderate person reliability and suboptimal item-person targeting indicate that further refinement is needed before its use for individual diagnostic assessment.

This is an open access article under the [CC BY-NC-SA](#) license.



Corresponding Author:

Thoiroh

Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi, Indonesia; thoi0108.2004@gmail.com

1. INTRODUCTION

Learning in Islamic Religious Education (PAI) still faces obstacles due to limited media and learning resources, as well as insufficient time allocation (Pratiwi and Maharani 2020). The delivery of material also

tends to be unfocused, as teachers' explanations often extend beyond the core material, leading to a lack of focus and inadequate development of students' analytical skills (Udin By Arifin et al., 2021). In addition, teachers' lack of innovation in delivering lessons causes students to tend to memorize rather than deeply understand the material (Fauzi & Inayati, 2023). PAI material also remains insufficiently contextual and fails to encourage students to connect Islamic values with real and social life (Fadkhurrohmad & Nursikin, 2023). even though its role is essential in shaping tolerance and Islamic character (Karim et al., 2020).

The assessment of higher-order thinking skills (HOTS) has become a global priority in 21st-century educational reform. Organizations such as UNESCO and the OECD, through the Program for International Student Assessment (PISA), consistently emphasize that educational success is not measured solely by rote memorization, but rather by students' capacity to analyze, evaluate, and create (OECD, 2023; Anwar, 2025). In the context of Islamic education, this demand aligns with the fundamental purpose of teaching faith and ethics, which is to shape Muslim individuals who do not merely "know" but are also capable of "reflecting" and internalizing divine values through concrete actions (Afikah et al., 2022; Asnawi & Mas'ad, 2025). At the national level, the Merdeka Curriculum and the National Education Standards also mandate the strengthening of critical and creative thinking skills as graduate competencies, which explicitly require assessment instruments capable of measuring higher-order cognitive domains (Kemendikbudristek, 2024).

Although the importance of HOTS has been widely recognized, the implementation of assessment in the Akidah Akhlak subject at madrasahs reveals significant gaps. Based on preliminary observations and a review of the literature, the issues in Akidah Akhlak instruction are not singular but are divided into three interrelated aspects that must be clearly distinguished. First, regarding teaching methods, teachers still rely heavily on lectures and make insufficient use of interactive media, resulting in a learning process that is more oriented toward the transfer of information than the development of moral reasoning (Fauzi & Inayati, 2023). Second, regarding instructional materials, the teaching resources used—such as textbooks and worksheets—tend to present content in a decontextualized manner, detached from students' social realities, and fail to link moral values to contemporary issues (Fadkhurrohmad & Nursikin, 2023; Susilawati et al., 2023). Third, and the primary focus of this study, is the assessment aspect. The assessment instruments used are still dominated by questions that measure memorization and basic understanding (cognitive levels C1–C2), while questions requiring analysis (C4), evaluation (C5), and creation (C6) are very rarely encountered (Pratiwi & Maharani, 2020). As a result, assessment results do not reflect students' critical and reflective thinking skills, and teachers also acknowledge their limited competence in designing valid and contextual HOTS items (Utami et al., 2025).

A research gap is evident from several previous findings. Studies by (Fauzi & Inayati, 2023) have identified that Islamic Education (PAI) test items still rely heavily on rote memorization; however, neither study has systematically developed assessment instruments based on a standardized development model. Meanwhile, the study by (Mufrihah et al., 2025) used Rasch modeling to measure a morality scale in Islamic boarding schools, but it did not specifically target the *Akidah Akhlak* subject and did not explicitly integrate HOTS content. As for ADDIE-based development research, such as that conducted by (Maxnun et al., 2024), it has focused more on digital learning media rather than on the development of assessment instruments. Thus, there has been no study that comprehensively develops and validates a HOTS assessment instrument for the *Akidah Akhlak* subject using modern psychometric approaches such as Rasch modeling, which can provide independent and objective estimates of student ability and item parameters, and is not dependent on a specific sample—unlike the limitations of Classical Test Theory (CTT), which is both sample-dependent and test-dependent (Boone, 2016; Bond et al., 2020).

Rasch modeling was chosen for this study because it allows for the analysis of model fit, item difficulty, and differential item functioning (DIF)—all of which are highly relevant for ensuring fairness and accuracy in the context of religious education, which is steeped in local cultural values.

Based on the above, this study aims to develop a HOTS-based assessment instrument for the *Akidah Akhlak* subject in madrasahs and to test its validity and reliability using Rasch modeling. The research

questions posed are: (1) What is the procedure for developing a HOTS instrument for *Akidah Akhlak* based on the ADDIE model? (2) What is the psychometric quality of the instrument's items as assessed through Rasch modeling? (3) What is the level of students' HOTS abilities as measured by this instrument?

The novelty of this study lies in three aspects. First, this study specifically developed a test blueprint and HOTS test items for the *Akidah Akhlak* subject—which had not previously been available—based on the revised Bloom's taxonomy and the contemporary Islamic context. Second, this study is the first to integrate the ADDIE development model with Rasch modeling within a single framework for validating religious education instruments, thereby yielding more comprehensive validity evidence. Third, this study produces an assessment instrument that is not only content-valid but also possesses objective and fair psychometric characteristics, making it widely applicable as an assessment model aligned with 21st-century demands. The significance of this study is not only theoretical—in enriching the field of Islamic education measurement—but also practical, as it provides a ready-to-use assessment tool for madrasah teachers to diagnose and develop students' higher-order thinking skills more accurately and meaningfully.

2. METHODS

2.1 Research Type and Design

This study is classified as research and development (R&D), aimed at producing a HOTS-based assessment instrument for the *Akidah Akhlak* subject, while also testing the empirical validity of the instrument (Sugiyono, 2020). Thus, this study is not merely a validation study of an existing instrument but encompasses the entire development cycle, from design to field testing. The design used is the ADDIE model (Analysis, Design, Development, Implementation, Evaluation) because this model is systematic, iterative, and has been proven effective in the development of educational instruments (Branch, 2010).

2.2 Participants and Sampling Method

This study involved 11th- and 12th-grade students at Madrasah Aliyah (MA) schools in Jambi Province. The selection of the 11th and 12th grades was based on the consideration that the *Akidah Akhlak* curriculum at these two levels covers complex theological and ethical themes, making them suitable for assessing higher-order thinking skills (Kemendikbudristek, 2024). The sampling technique used purposive sampling with the following criteria: (a) the madrasah had implemented the Merdeka Curriculum or the 2013 Curriculum, (b) the *Akidah Akhlak* teacher at the madrasah expressed willingness to participate, and (c) the madrasah had adequate access to a computer lab for administering digital-based tests. Based on these criteria, three public madrasahs and two private madrasahs were selected, spread across three regencies/cities in Jambi, representing both urban and rural areas.

The participants consisted of two groups. The first group comprised 30 students for small-scale pilot testing, which aimed to identify technical weaknesses in the test items (e.g., ambiguous wording, layout errors, or time constraints) prior to large-scale testing. The sample size of 30 students was selected based on the minimum recommendation for initial testing in instrument development (Johanson & Brooks, 2010). The second group consisted of 300 students for large-scale field testing, which aimed to empirically estimate item parameters and reliability. The sample size of 300 respondents met the requirements for Rasch analysis, which requires a minimum of 200–300 respondents to obtain stable parameter estimates (Johanson & Brooks, 2010; Kunselman, 2024). The demographics of the participants included 48% male students and 52% female students, with an age range of 16–18 years. Prior to data collection, all participants and their parents or guardians were provided with a written explanation of the study's objectives and signed informed consent forms. Furthermore, this study obtained ethical approval from the institutional ethics committee and official permission from the madrasah principal and the Jambi Provincial Education Office.

2.3 Instrument Development and Blueprint

The instrument developed is a multiple-choice test with four answer options (distractors), consisting of 40 items. The development followed the ADDIE procedure as follows:

a. Analysis Phase (Analysis)

A needs analysis was conducted through a review of the syllabus and lesson plans for the Akidah Akhlak courses in grades 11 and 12, as well as interviews with five subject teachers to identify the core competencies (KD) that allow for the measurement of Higher-Order Thinking Skills (HOTS). The results of the analysis showed that the core competencies in the chapters on Belief in the Last Day, Praiseworthy and Blameworthy Morals, and the Concepts of Kalam and Qadar had never been assessed with questions requiring analysis and evaluation.

b. Design Phase (Design)

An instrument blueprint was developed covering three content domains of Akidah Akhlak: (a) Islamic Theology (Tawhid and Asma'ul Husna), (b) Ethics (praiseworthy and blameworthy behavior), and (c) Concepts of Divinity (Qada' and Qadar). Each domain was developed into HOTS indicators based on the revised Bloom's taxonomy (Anderson et al., 2023), with the following distribution: C4 (Analyzing) – 18 items (45%), measures students' ability to break down ethical concepts into their constituent elements and identify cause-and-effect relationships in moral cases. C5 (Evaluating) – 14 items (35%), measures students' ability to assess the validity of a theological statement based on naqli and aqli arguments, as well as to weigh moral decisions in dilemmatic situations. C6 (Creating) – 8 items (20%), measures students' ability to formulate solutions or alternative courses of action based on Islamic values in a contemporary social context.

Each test item is accompanied by a stimulus in the form of a real-life case, a hadith, a verse from the Qur'an, or a situational narrative to align with the principle of contextual assessment (Susilawati et al., 2023).

c. Development Stage

The test items that had been designed were then validated by five experts, consisting of: two Islamic education professors specializing in learning assessment, two senior *Akidah Akhlak* teachers (each with >10 years of experience), and one educational measurement expert. The experts were asked to evaluate each item using a validation sheet with a 1–4 Likert scale across three aspects: (a) alignment with the indicators, (b) alignment with cognitive levels C4/C5/C6, and (c) clarity of language and stimuli. The expert validation data were analyzed using Aiken's V index, with the following interpretation criteria: $V \geq 0.80$ is considered to have high content validity, $V = 0.60 - 0.79$ is categorized as adequate, and $V < 0.60$ is revised or discarded (Alghadir et al., 2018). Items with a $V < 0.60$ were revised and revalidated until they met the threshold.

d. Implementation Stage

A small-scale pilot test was conducted with 30 students to assess readability, completion time, and clarity of the stimuli. Based on the results of this pilot test, some question wording was revised, and two items were replaced because their readability was too low. Subsequently, a large-scale pilot test was conducted with 300 students under standardized test conditions, with a time limit of 90 minutes. The test was administered digitally using the Google Forms platform to facilitate data collection and avoid data entry mistakes.

e. Evaluation Phase

Data from the large-scale pilot test were analyzed quantitatively using Rasch modeling to estimate empirical validity, reliability, and item fit.

2.4 Data Analysis Using Rasch Modeling

Data analysis in this study utilized Rasch modeling with a dichotomous model (1-PL) because the data collected were dichotomous (correct answer = 1, incorrect = 0). The Rasch model was chosen because it has advantages over Classical Test Theory (CTT), namely: (a) it produces estimates of student ability and item difficulty that are sample-independent and test-independent, (b) it can present item fit information separately, and (c) it can detect item bias or differential item functioning (DIF) more objectively (Boone, 2016; Bond et al., 2020). The analysis was conducted using the QUEST software (Adams & Khoo, 1996), which is specifically designed for the dichotomous Rasch model. The parameters analyzed include:

a. Item Fit

Item fit was assessed using the Infit Mean Square (MNSQ) and Outfit Mean Square (MNSQ) values. The decision criteria followed the recommendations of (Linacre, 2002; Dong et al., 2025):

- An item is considered to fit if the Infit and Outfit MNSQ values fall within the range of 0.77–1.30.
- If the MNSQ value is < 0.77 , the item is considered a misfit because it is overfit and lacks discriminative power.
- If the MNSQ value is > 1.30 , the item is considered a misfit because it is underfit and does not contribute to the measurement.
- Items that are misfits will be eliminated or revised before being used in the final stage.

b. Reliability

Reliability was estimated using the Person Separation Reliability (PSR) and Item Separation Reliability (ISR) indices. A $PSR \geq 0.80$ indicates that the instrument is able to consistently distinguish between students with different abilities, while an $ISR \geq 0.90$ indicates that the test items have a well-distributed difficulty level (Caselli et al., 2025).

c. Item Difficulty Level (Item Calibration)

Item difficulty is expressed in logit units, with higher values indicating more difficult items. Difficulty levels are classified according to the following guidelines: 1) Logit value < -1.0 : very easy item. 2) $-1.0 \leq \text{logit} \leq 1.0$: moderate item. 3) Logit value > 1.0 : very difficult item (Bond et al., 2020).

The entire analysis process was conducted transparently and documented to ensure the replicability of the research. Raw data, analysis scripts, and QUEST output results are stored in a research archive that is accessible to other researchers upon valid request.

3. FINDINGS AND DISCUSSION

3.1 Content Validity (Expert Judgment)

The initial process to ensure the quality of the instrument was conducted through content validation by five experts (*expert judgment*), consisting of three lecturers from UIN STS Jambi, one senior Islamic Education teacher, and one peer reviewer. The assessment was carried out on 40 multiple-choice items based on HOTS in the Akidah Akhlak material. Inter-rater consistency is presented in Table 1.

Table 1. Inter-Rater Assessment Consistency

Score	R1	R2	R3	R4	R5
Mean Score	3.78	3.78	3.76	3.85	3.83
Standard Deviation	0.13	0.12	0.12	0.15	0.12

The mean score per rater ranged from 3.76 to 3.85 (on a 1–4 scale), with standard deviations between 0.12 and 0.15. These values indicate a high level of agreement among the raters and minimal ambiguity in item construction.

Subsequently, the expert assessment data were analyzed quantitatively using Aiken's V index. The results of the analysis for the 40 items are presented in Table 2.

Table 2. Summary of Aiken's V Index Analysis per Item

Aiken's V Coefficient Range	Number of Items	Classification
0.83 – 0.89	3 items	Valid
0.90 – 0.97	35 items	Valid
0.98 – 1.00	2 items	Valid
Mean	0.93	Very Valid

The Aiken's V coefficients ranged from 0.83 to 1.00, with an overall mean of 0.93. All 40 items exceeded the minimum validity threshold ($V \geq 0.80$), thus categorized as content-valid. However, high content validity does not guarantee empirical functioning of the items; thus, further testing through the Rasch model was required.

3.2 Item Parameter Analysis Using the Rasch Model

After content validity was established, empirical analysis was conducted using the Rasch model with the assistance of the QUEST software. The analysis covered three aspects: item fit, reliability, and item difficulty. Of the 40 items tested, one item (Item 1) was identified as misfit and was excluded from subsequent analyses. Consequently, the final instrument consisted of 39 items used for parameter estimation and reliability analysis.

3.3 Item Fit Analysis

Item fit analysis employed the Infit Mean Square (MNSQ), Outfit MNSQ, and their accompanying t-statistics. The summary statistics for the 39 retained items are presented in Table 3.

Table 3. Rasch Model Item Fit Statistics (N = 39 items)

Statistic	Mean	Standard Deviation
Infit MNSQ	0.98	0.11
Outfit MNSQ	0.99	0.38
Infit t	0.1	1.2
Outfit t	0.1	1.6

The mean Infit MNSQ (0.98) and Outfit MNSQ (0.99) values were close to the ideal value of 1.00, indicating good fit between the data and the model overall. Based on the fit criteria ($\text{Outfit } t \leq 2.00$), all 39 items (100%) were categorized as fit. No misfit items remained after the removal of Item 1.

Table 4. Summary of Item Fit Based on Outfit t (N = 39 items)

Category	Number of Items	Item Numbers
Fit ($\text{Outfit } t \leq 2.00$)	39 items	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
Misfit ($\text{Outfit } t > 2.00$)	0 items	-

3.4 Reasons for Removing Item 1

Item 1 had an Outfit t value of 2.3 in the initial analysis, which exceeded the threshold of 2.00. This underfit condition indicated that students' response patterns to this item were more random than predicted by the model, which could be attributed to: (a) ambiguous wording allowing multiple interpretations, (b) unclear stimulus, or (c) the presence of more than one seemingly correct answer choice. Based on the consideration that revising the item individually did not guarantee significant quality improvement, and to maintain the integrity of the psychometric analysis, Item 1 was removed from the final instrument. This

decision aligns with common practice in Rasch-based instrument development, where misfit items that cannot be substantially improved are better removed than retained (Bond et al., 2020; Linacre, 2002).

3.5 Reliability

Instrument reliability was estimated through two indicators: item reliability and person reliability. The estimation results for the 39 items are presented in Table 5.

Table 5. Rasch Model Reliability Estimation Results (N = 39 items)

Estimation Type	Reliability Value	Category
Item Reliability	0.97	High
Person Reliability	0.69	Moderate

An item reliability of 0.97 indicates that the ordering of item difficulty levels is highly stable and can be replicated in other samples. Person reliability of 0.69 falls into the moderate category, meaning that the instrument has limited consistency in distinguishing individual students' ability levels. This value showed a slight increase from the previous 0.68, indicating that the removal of Item 1 provided a small improvement in measurement consistency. However, this value still suggests that there remains response variation that cannot be fully explained by differences in students' cognitive abilities.

3.6 Item Difficulty and Item-Person Targeting

Item difficulty analysis was conducted by calibrating each item on the logit scale. The distribution summary for the 39 items is presented in Table 6.

Table 6. Distribution of Item Difficulty Levels Based on Logit Values (N = 39 items)

Difficulty Level	Logit Range	Number of Items	Item Numbers
Very Difficult	> +2.0	2	26, 39
Difficult	+1.0 to +2.0	6	15, 23, 24, 25, 29, 30
Moderate	0.0 to +0.9	11	4, 10, 11, 13, 22, 27, 28, 32, 33, 34, 38
Easy	-1.0 to -0.1	16	3, 5, 7, 8, 9, 12, 14, 16, 17, 18, 19, 20, 21, 31, 35, 36
Very Easy	< -1.0	4	2, 6, 37, 40

The mean respondent ability was at logit 2.05, while the mean item difficulty was at logit 0.00 (the Rasch model standard). The difference of 2.05 logits indicates that, on average, the items in this instrument tended to be easier than the average ability of the students. In other words, the test was not too difficult for the sample tested, and most students were able to answer most items correctly. This finding suggests that the instrument still requires the addition of items with higher difficulty levels to reach students with above-average HOTS abilities.

Discussion

The findings demonstrate that integrating the ADDIE development model with Rasch analysis provides a systematic approach to developing a Higher-Order Thinking Skills (HOTS) assessment instrument for *Akidah Akhlak* learning. The development process resulted in 39 retained multiple-choice items following expert validation and empirical testing. This approach addresses the need for assessment instruments that extend beyond memorization and basic comprehension toward analytical, evaluative, and creative thinking. The combination of instructional design procedures and psychometric evaluation is consistent with Branch's (2010) framework for systematic educational development and Boone's (2016) emphasis on the importance of Rasch analysis in establishing measurement quality. Nevertheless, the instrument's practical usefulness must be interpreted in relation to its reported reliability and targeting limitations.

The mean Aiken's V coefficient of 0.93 indicates strong agreement regarding the content relevance of the developed items. Expert evaluation focused on alignment with learning indicators, HOTS cognitive levels, and clarity of language and stimuli. These findings suggest that the instrument adequately

represents the intended *Akidah Akhlak* content according to the participating validators. Content validation is particularly important because HOTS assessment requires more than difficult questions; items must elicit reasoning that corresponds to the intended cognitive processes. Brookhart (2010) emphasizes the importance of aligning assessment tasks with higher-order cognitive demands. However, content validity alone cannot establish whether items function appropriately when administered to students. Consequently, empirical analysis remains necessary to complement expert judgment.

Rasch analysis identified one misfitting item among the original 40 questions, resulting in a final instrument of 39 items. The excluded item demonstrated an Outfit t statistic of 2.3, exceeding the criterion applied in the reported analysis. The remaining items satisfied the reported Outfit t threshold of ≤ 2.00 , while their mean Infit and Outfit mean-square statistics were 0.98 and 0.99, respectively. These results suggest satisfactory overall correspondence between observed responses and model expectations. As Bond et al. (2020) explain, Rasch fit statistics provide evidence about the extent to which response patterns conform to the measurement model. Nevertheless, the reported aggregate statistics should be supplemented by item-level fit information before concluding that every retained item satisfies all prespecified fit criteria.

The reliability findings reveal an important distinction between item calibration and the precision of student ability measurement. Item reliability reached 0.97, indicating that the relative ordering of item difficulty was estimated consistently within the study sample. In contrast, person reliability was 0.69, suggesting limited precision in differentiating students with similar HOTS abilities. This distinction is important because strong item reliability does not automatically establish equally strong measurement of individual respondents. Bond et al. (2020) emphasize the need to interpret reliability alongside the distribution of person abilities and item difficulties. Therefore, the instrument appears more suitable for preliminary classroom-level assessment than for high-stakes decisions or precise individual diagnosis.

The item–person targeting results provide further insight into the moderate person reliability. The mean student ability was 2.05 logits, whereas the mean item difficulty was centered at 0.00 logits. This difference indicates that the test items were generally easier than the ability level of the participating students. Furthermore, 20 of the 39 retained items were classified as easy or very easy, while only two were categorized as very difficult. Such an imbalance may reduce measurement precision among higher-performing students because the instrument provides relatively few challenging items capable of distinguishing their abilities. Rasch-based instrument development requires attention not only to model fit but also to the alignment between item difficulty and respondent ability (Boone, 2016; Bond et al., 2020). Future revisions should therefore prioritize additional challenging items, particularly those requiring complex evaluation of moral dilemmas and theological arguments.

Another important consideration concerns the measurement of different HOTS dimensions through multiple-choice questions. The original instrument included items targeting analyzing (C4), evaluating (C5), and creating (C6). Although contextual multiple-choice items can assess aspects of analysis and evaluation, their capacity to capture authentic creative performance is more limited because students select predetermined responses rather than independently construct solutions. Brookhart (2010) highlights the importance of assessment formats that match the thinking processes being evaluated. Accordingly, the C6 items should be interpreted cautiously, and future instrument development could incorporate constructed-response questions, case analyses, or performance-based tasks. Such complementary formats would provide broader evidence of students' ability to formulate original, contextually appropriate responses to ethical problems.

In relation to previous research, this study contributes a subject-specific instrument that integrates HOTS-oriented item development with empirical calibration. Earlier work has highlighted challenges in implementing HOTS-oriented Islamic education assessment (Fauzi & Inayati, 2023; Pratiwi & Maharani, 2020), while Maxnun et al. (2024) examined the development of HOTS-based cognitive assessment instruments using ADDIE. The present study extends this discussion by applying Rasch analysis to *Akidah*

Akhlak assessment and reporting item fit, reliability, and difficulty estimates. Its contribution should therefore be understood as a contextual and methodological application rather than an entirely unprecedented combination of ADDIE and Rasch approaches.

Overall, the findings indicate that the developed instrument has promising content validity and item-level measurement properties, but further refinement is required to strengthen its diagnostic precision. The moderate person reliability, relatively easy item distribution, and limitations of the multiple-choice format constrain the interpretation of individual HOTS achievement. In addition, testing was limited to selected madrasahs in Jambi Province, which restricts generalization to other educational settings. Subsequent research should evaluate the revised instrument across more diverse student populations, examine item functioning across relevant groups, and incorporate assessment formats that capture complex reasoning more directly. These improvements would support the development of a more comprehensive and defensible HOTS assessment system for Akidah Akhlak learning.

4. CONCLUSION

This study successfully developed a Higher-Order Thinking Skills (HOTS) assessment instrument for *Akidah Akhlak* learning by integrating the ADDIE development model with Rasch analysis. The final instrument consisted of 39 multiple-choice items after one misfitting item was removed from the initial 40 items. The findings demonstrated strong content validity (mean Aiken's $V = 0.93$), satisfactory item fit based on the reported Outfit t criterion (≤ 2.00), and high item reliability (0.97). However, person reliability was relatively moderate (0.69), indicating limited precision in distinguishing students with similar HOTS abilities. Furthermore, the mean student ability (2.05 logits) exceeded the mean item difficulty (0.00 logits), suggesting that the instrument was relatively easy for the participating students. These findings indicate that the developed instrument has potential as a classroom assessment tool for evaluating HOTS in *Akidah Akhlak* learning, although its use for individual diagnostic purposes requires caution. This study has several limitations, including its reliance on multiple-choice questions, which may not adequately capture authentic creative thinking, its restricted geographical coverage involving selected madrasahs in Jambi Province, and the suboptimal alignment between item difficulty and student ability. Future research should refine the instrument by incorporating more challenging items, particularly at the evaluating (C5) and creating (C6) levels, and integrating constructed-response questions or performance-based assessments. Further validation involving larger and more diverse student populations, including differential item functioning analysis, is also recommended to strengthen measurement precision, fairness, and generalizability. These improvements are expected to support the development of a more comprehensive, reliable, and contextually relevant HOTS assessment instrument for Islamic education.

Acknowledgment: The researcher expresses sincere appreciation and gratitude to the Lembaga Pengelola Dana Pendidikan (LPDP) and the Ministry of Religious Affairs (Program MORA THE AIR FUND) for their financial support, which served as the main pillar for this study's implementation. Deep thanks are also extended to Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi for providing the academic environment and facilities necessary for this research. Additionally, the researcher is grateful for the cooperation and assistance of the participating schools during the field data collection process, which was crucial to the completion of this scientific work

REFERENCES

- Afikah, A., Astuti, S. R. D., Suyanta, S., Jumadi, J., & Rohaeti, E. (2022). Mobile Learning in Science Education to Improve Higher-Order Thinking Skills (HOTS) and Communication Skills: A Systematic Review. *International Journal of Advanced Computer Science and Applications*, 13(7). <https://doi.org/10.14569/IJACSA.2022.0130782>
- Alghadir, A. H., Anwer, S., Iqbal, A., & Iqbal, Z. A. (2018). Test-retest reliability, validity, and minimum

- detectable change of visual analog, numerical rating, and verbal rating scales for measurement of osteoarthritic knee pain. *Journal of Pain Research*, 11. <https://doi.org/10.2147/JPR.S158847>
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Rath, J., & Wittrock, M. C. (2023). Revisi atas Taksonomi Bloom (et.al.) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives (Abridge Edition). *TAKSONOMI ANDERSON (et.AL.): Revisi Atas Taksonomi Bloom (et.AL.)*.
- Anwar, N. (2025). Report review: Reimagining our futures together: a new social contract for education. *Globalisation, Societies and Education*, 23(5). <https://doi.org/10.1080/14767724.2023.2269103>
- Asnawi, M. F. A., & Mas'ad, Z. (2025). Effectiveness of Islamic Education Learning Model Based on Higher Order Thinking Skills (HOTS): Systematic Review. *Progressive Journal of Educational Science*, 1(2). <https://doi.org/10.65729/progressive.v1i2.330>
- Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the rasch model: Fundamental measurement in the human sciences. In *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. <https://doi.org/10.4324/9780429030499>
- Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, 15(4). <https://doi.org/10.1187/cbe.16-04-0148>
- Branch, R. M. (2010). Instructional design: The ADDIE approach. In *Instructional Design: The ADDIE Approach*. <https://doi.org/10.1007/978-0-387-09506-6>
- Brookhart, S. M. (2010). *How to Assess Higher-Order Thinking Skills in Your Classroom*. In *Ascd*.
- Caselli, S., Leonardi, M., Magnani, F. G., Cacciatore, M., Barbadoro, F., Ippoliti, C., Kreiner, S., Pellicciari, L., & La Porta, F. (2025). Comparing the Different Sets of Item-Level Diagnostic Criteria of the Coma Recovery Scale-Revised (CRS-R): A Measurement-Based Approach Driven by Rasch Analysis. *Archives of Physical Medicine and Rehabilitation*, 106(7). <https://doi.org/10.1016/j.apmr.2024.12.009>
- Dong, Y., Xu, W., Huang, J., & Yann, K. (2025). Validating and refining a multi-dimensional scale for measuring AI literacy in education using the Rasch Model. *Humanities and Social Sciences Communications*, 12(1). <https://doi.org/10.1057/s41599-025-05670-6>
- Fadkhurrohmad, M. N., & Nursikin, M. (2023). Pendekatan dalam inovasi materi dan metode pembelajaran Kurikulum PAI berbasis HOTS Era Industri 4 . 0 di MTs NU Salatiga. *Jurnal Pengabdian Kepada Masyarakat Nusantara (JPkMN)*, 4(1), 386–395.
- Fauzi, A., & Inayati, N. L. (2023). Implementasi Evaluasi Pembelajaran Pendidikan Al Islam di Sekolah Menengah Atas Muhammadiyah. *Munaddhomah*, 4(2), 272–283. <https://doi.org/10.31538/munaddhomah.v4i2.438>
- Helmiyatun, Taqiudin, M., & Ahmad. (2025). Content Validity of Student Perception Instrument on School Educational Management. *Journal of Practice Learning and Educational Development*, 5(1). <https://doi.org/10.58737/jpled.v5i1.422>
- Johanson, G. A., & Brooks, G. P. (2010). Initial scale development: Sample size for pilot studies. *Educational and Psychological Measurement*, 70(3). <https://doi.org/10.1177/0013164409355692>
- Karim, R., Mandang, F. H., Luntungan, G., & Rotty, V. N. J. (2020). Improving Competency Hosted Skills through Think-Pair-Share Cooperative Learning Models in Class XI Science Students of SMA Negeri 1 Rataan. *Asia Pacific Journal of Management and Education*, 3(3), 32–36. <https://doi.org/10.32535/apjme.v3i3.965>
- Kemendikbudristek. (2024). Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 8 Tahun 2024 Tentang Standar Isi Pada PAUD, Jenjang pendidikan Dasar dan Jenjang Pendidikan Menengah. In *Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi*.
- Kunselman, A. R. (2024). A brief overview of pilot studies and their sample size justification. In *Fertility and Sterility* (Vol. 121, Issue 6). <https://doi.org/10.1016/j.fertnstert.2024.01.040>
- Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? Rasch measurement transaction. In <https://www.rasch.org/rmt/rmt162.pdf> (Vol. 16, Issue 2).

- Maxnun, L., Kristiani, K., & Sulistyaningrum, C. D. (2024). Development of hots-based cognitive assessment instruments: ADDIE model. *Journal of Education and Learning*, 18(2). <https://doi.org/10.11591/edulearn.v18i2.21079>
- Mufrihah, A., Yudha, E. S., Supriatna, M., Ahman, A., & Nurihsan, J. (2025). Rasch Model Analysis of Santri Reverence Morals Scale. *Islamic Guidance and Counseling Journal*, 8(1), 1–18. <https://doi.org/10.25217/0020258548700>
- Ocy, D. R., Rahayu, W., & Arthur, R. (2024). Rasch model analysis: Evaluating students' spatial thinking ability in higher-order thinking skills trigonometric comparison assessment (HOTS-TCA). *Jurnal Gantang*, 9(2). <https://doi.org/10.31629/jg.v9i2.7386>
- OECD. (2023). PISA 2022 Assessment and Analytical Framework. In *OECD Publishing*.
- Pratiwi, Z. I., & Maharani, D. (2020). Penerapan Pembelajaran Pendidikan Agama Islam (Pai) Berbasis Higher Order Thinking Skills (HOTS) (Studi Analisis Pada Kelas XI di SMA Dharma Karya UT Tangerang Selatan). *Jurnal Qiro'ah*, 10(2), 57–72.
- Rohman, M. H., Marwoto, P., Nugroho, S. E., & Supriyadi, S. (2024). Effectiveness of Ethnoecological-STEM Project-Based Learning Model to Improve Critical Thinking Skills, Creativity, and Science Concept Mastery. *International Journal of Cognitive Research in Science, Engineering and Education*, 12(3). <https://doi.org/10.23947/2334-8496-2024-12-3-521-534>
- Sugiyono. (2020). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D*. ALFABETA.
- Susilawati, S., Yaqin, M. Z. N., Wahidmurni, Ulfa, F. A., & Krismoneta. (2023). Training on Preparation of Islamic Integrated HOTS-Based Questions and Their Application in Online Learning in Elementary Schools. *Munaddhomah*, 4(4), 867–877. <https://doi.org/10.31538/munaddhomah.v4i4.459>
- Udin By Arifin, M. B., Nurdyansyah, Rindaningsih, I., & Kalimah, S. (2021). Development of Smart Play Wheel Learning Media to Improve Student Learning Outcomes in Islamic Elementary schools. *Journal of Physics: Conference Series*, 1779(1). <https://doi.org/10.1088/1742-6596/1779/1/012049>
- Utami, W. S., Kusaeri, Ridho, A., Yusuf, A., Widodo, B. S., Prastiyono, H., & Murtini, S. (2025). Unraveling social and economic problems: what basics of critical thinking skills are needed? *International Journal of Evaluation and Research in Education*, 14(1), 566–574. <https://doi.org/10.11591/ijere.v14i1.29180>