

Development and psychometric evaluation of a formative assessment instrument for higher order thinking skills in Akidah Akhlak Phase E

Chamelia Bunga Nurani^{1*}, Rizki Rahmawati²

^{1,2} UIN Sayyid Ali Rahmatullah, Tulungagung, Indonesia

*Corresponding Author: chameliabunganurani@gmail.com

Submission: 26 June 2026

Revision: 10 July 2026

Accepted: 16 July 2026

Abstract

Formative assessment plays an essential role in fostering *Higher Order Thinking Skills* (HOTS); however, the development of HOTS assessment instruments for Akidah Akhlak supported by comprehensive psychometric evaluation remains limited. This study aimed to develop a formative HOTS assessment instrument for Akidah Akhlak Phase E and evaluate its psychometric quality. An *instrument development study* was employed through construct analysis, instrument development, content validation by two experts, field testing involving 35 students, and psychometric evaluation using Aiken's V, *Product Moment* correlation, *Cronbach's Alpha*, and ITEMAN. The findings revealed a mean Aiken's V of 0.846, 34 of 35 items (97.14%) met the empirical validity criteria, a reliability coefficient of 0.929, and item characteristics were dominated by moderate difficulty, very good discrimination, and fully functioning distractors. This study contributes a valid and reliable HOTS assessment instrument and a psychometric-based instrument development framework to support *evidence-based formative assessment* in Akidah Akhlak learning.

Keywords: Formative assessment; *Higher Order Thinking Skills*; Instrument development; Psychometric evaluation; Akidah Akhlak.

Abstrak

Asesmen formatif berperan penting dalam mengembangkan *Higher Order Thinking Skills* (HOTS), namun pengembangan instrumen pada mata pelajaran Akidah Akhlak yang disertai evaluasi psikometrik secara komprehensif masih terbatas. Penelitian ini bertujuan mengembangkan instrumen asesmen formatif HOTS pada mata pelajaran Akidah Akhlak Fase E serta mengevaluasi kualitas psikometriknya. Penelitian menggunakan pendekatan *instrument development study* melalui analisis konstruk, pengembangan instrumen, validasi isi oleh dua ahli, uji coba kepada 35 peserta didik, serta evaluasi psikometrik menggunakan Aiken's V, korelasi *Product Moment*, *Cronbach's Alpha*, dan ITEMAN. Hasil penelitian menunjukkan nilai rata-rata Aiken's V sebesar 0,846, sebanyak 34 dari 35 butir (97,14%) memenuhi validitas empiris, reliabilitas sebesar 0,929, serta karakteristik butir didominasi tingkat kesukaran sedang, daya beda sangat baik, dan distraktor yang seluruhnya berfungsi efektif. Penelitian ini berkontribusi menyediakan instrumen HOTS yang valid, reliabel, serta model pengembangan instrumen berbasis evaluasi psikometrik untuk mendukung *evidence-based formative assessment* pada pembelajaran Akidah Akhlak.

Kata Kunci: Asesmen formatif; *Higher Order Thinking Skills*; Pengembangan instrumen; Evaluasi psikometrik; Akidah Akhlak

Introduction

The growing emphasis on formative assessment indicates that educational assessment is no longer viewed merely as a means of measuring learning outcomes but as an integral component of the learning process itself (Leenknecht et al., 2021). The shift from *assessment of learning* to *assessment for learning* positions assessment as a source of information for monitoring students' competency development, providing feedback, and continuously improving instructional practices (Al-Htaybat et al., 2018). Within this paradigm, the quality of the information generated depends largely on the quality of the assessment instrument employed (Gärtner et al., 2018). Instruments that fail to accurately represent the intended competencies may lead to inaccurate interpretations, resulting in less effective instructional decisions. Therefore, developing assessment instruments that meet established measurement standards has become essential for supporting learning processes that promote both instructional quality and student achievement.

This paradigm shift has become increasingly relevant with the implementation of the *Kurikulum Merdeka*, which positions formative assessment as an integral part of the learning process to monitor the attainment of learning objectives while fostering *Higher Order Thinking Skills* (HOTS) (Safrudin, 2025). Students are expected not only to recall information but also to analyze, evaluate, and generate solutions to various problems through cognitive processes at the C4, C5, and C6 levels of the Revised Bloom's Taxonomy (Arlianty et al., 2018). However, these objectives are difficult to achieve when the assessment instruments employed are unable to accurately measure higher-order thinking skills. This challenge is reflected in the 2022 *Programme for International Student Assessment* (PISA), in which Indonesia scored 359 in reading literacy, 366 in mathematics, and 383 in science, all of which remain below the OECD average (Solihin et al., 2024). These findings indicate that developing students' higher-order thinking skills remains a significant educational challenge that requires high-quality assessment instruments.

Addressing this challenge requires more than simply designing HOTS-oriented test items; it also demands assessment instruments with sound psychometric properties to ensure that assessment results can serve as a trustworthy basis for instructional decision-making (Schuwirth & van der Vleuten, 2018). A high-quality instrument should demonstrate validity, reliability, and appropriate item characteristics, including difficulty level, discrimination index, and distractor effectiveness (Friatma & Anhar, 2019). Without comprehensive psychometric evaluation, assessment instruments may produce biased information, preventing an accurate representation of students' abilities (Landers et al., 2022). Therefore, formative assessment instruments should be developed systematically through instrument specification, HOTS-based item construction, expert validation, and empirical testing before being

implemented in educational practice. Such a systematic approach ensures that the instrument is not only conceptually sound but also supported by robust empirical evidence for *evidence-based assessment*.

Studies on the development of *Higher Order Thinking Skills* assessment instruments have been widely conducted across various disciplines, including Islamic Education. Research by Sugiyono et al. (2018) demonstrated that the developed instruments generally met validity and reliability standards, making them suitable for instructional use. Similarly, Isnaeni et al. (2021) found that HOTS assessment instruments enabled teachers to identify students' critical thinking abilities more effectively than conventional assessments. Nevertheless, most existing studies have focused primarily on instrument development or reported validity and reliability as the principal indicators of instrument quality. Comprehensive psychometric evaluation, encompassing item difficulty, discrimination, and distractor effectiveness, remains relatively limited, despite the critical role these characteristics play in ensuring that each test item accurately and consistently measures the intended competencies.

Based on these considerations, this study aims to develop a formative assessment instrument for *Higher Order Thinking Skills* in Akidah Akhlak Phase E and comprehensively evaluate its psychometric quality. The instrument was developed through instrument specification, HOTS-based item construction, expert content validation, and empirical testing to obtain evidence of validity, reliability, item difficulty, discrimination, and distractor effectiveness. Unlike previous studies that primarily emphasized either instrument development or partial psychometric evaluation, this study integrates both aspects within a comprehensive evaluation framework. Consequently, the study not only produces a formative assessment instrument suitable for Akidah Akhlak Phase E learning but also proposes a psychometric-based instrument development model that supports *evidence-based assessment* practices and serves as a reference for developing HOTS assessment instruments in broader educational contexts.

Method

This study employed an *instrument development study* design to develop a formative assessment instrument for *Higher Order Thinking Skills* (HOTS) in Akidah Akhlak Phase E and evaluate its psychometric quality (Merle, 1998). This approach focuses on developing educational measurement instruments through instrument specification, item construction, validation, field testing, and psychometric evaluation to produce an instrument that meets established educational measurement standards. The development process consisted of analyzing the Learning Outcomes (*Capaian Pembelajaran* [CP]), Learning Objectives (*Tujuan Pembelajaran* [TP]), and HOTS indicators at the C4–C6 cognitive levels of the Revised Bloom's Taxonomy; developing the blueprint, test specification, multiple-choice items, answer keys, and scoring

guidelines; conducting content validation by two experts specializing in educational assessment and Akidah Akhlak; administering a field test to 35 Phase E students at MAN 2 Blitar; and revising the instrument based on the psychometric evaluation results. The initial instrument comprised 35 HOTS-based multiple-choice items.

The research data consisted of qualitative data in the form of expert evaluations, comments, and suggestions from the validators, as well as quantitative data derived from students' responses during the field test. Content validity was examined using Aiken's V coefficient, whereas empirical validity was assessed using the Pearson Product-Moment correlation between individual item scores and the total test score (Aiken, 1985). Reliability was estimated using Cronbach's Alpha coefficient, while item characteristics were analyzed using the ITEMAN software package, including the *difficulty index*, *discrimination index*, and *distractor effectiveness* (Astin & Antonio, 2012). The findings from these analyses served as the basis for revising the instrument, resulting in a final set of 34 items that satisfied the empirical validity criteria and demonstrated adequate psychometric quality for use as a formative HOTS assessment instrument in Akidah Akhlak Phase E learning.

Results and Discussion

Development of a formative assessment instrument for higher order thinking skills in Akidah Akhlak Phase E

The development of the *Higher Order Thinking Skills* (HOTS) formative assessment instrument in this study aimed to produce a valid and systematic measurement tool for assessing students' higher-order thinking skills in Akidah Akhlak Phase E. The instrument was developed based on the Learning Outcomes (*Capaian Pembelajaran* [CP]) and Learning Objectives (*Tujuan Pembelajaran* [TP]) of the *Kurikulum Merdeka* to ensure that each test item accurately represented the intended competencies. An analysis of the learning content indicated that students were expected not only to understand Islamic creed concepts conceptually but also to analyze phenomena, evaluate problems, and determine appropriate solutions based on Islamic values. Accordingly, the instrument was designed to assess cognitive processes at the C4 (*analyzing*), C5 (*evaluating*), and C6 (*creating*) levels of the Revised Bloom's Taxonomy, thereby supporting the implementation of formative assessment focused on competency development.

Based on this analysis, the development process began with the preparation of an instrument blueprint that established the alignment among the Learning Outcomes, Learning Objectives, item indicators, cognitive levels, instrument format, and the number of items to be developed. The blueprint subsequently served as the basis for constructing the table of specifications, ensuring that each indicator was proportionally represented during item writing. A total of 35 multiple-choice items were then developed using contextual stimuli, including narratives, case illustrations,

and everyday-life situations relevant to the Akidah Akhlak curriculum. These contextual stimuli were intended to encourage students not only to recall concepts but also to apply their knowledge, engage in reasoning, and make decisions grounded in the values of *akidah* and *akhlak* across various learning contexts.

Table 1. Specification of the developed HOTS formative assessment instrument

Component	Specification
Subject	Akidah Akhlak
Phase	E
Learning Material	Allah's Necessary Attributes, Impossible Attributes, and Permissible Attributes
Instrument Type	Multiple-choice formative assessment
Initial Number of Items	35
Cognitive Levels	C4 (<i>Analyzing</i>), C5 (<i>Evaluating</i>), C6 (<i>Creating</i>)
Basis for Development	Learning Outcomes (CP) and Learning Objectives (TP) of Phase E

Source: Developed by the Author (2026).

As presented in Table 1, the instrument was specifically developed for the topic of *Allah's Necessary Attributes, Impossible Attributes, and Permissible Attributes* in Phase E, with a particular emphasis on measuring higher-order thinking skills. Each test item was designed to assess a single competency indicator, thereby maintaining a clear alignment among the indicators, learning objectives, and the HOTS construct. In addition to ensuring content relevance, the item development process followed established principles of multiple-choice test construction, including clear item stems, homogeneous answer options, communicative language, and plausible distractors. These characteristics were expected to produce an assessment instrument capable of providing comprehensive information regarding students' abilities to analyze, evaluate, and solve problems related to Akidah Akhlak learning.

The development stage resulted in an initial assessment package consisting of the instrument blueprint, the table of specifications, 35 multiple-choice items, an answer key, and scoring guidelines, all of which were prepared according to the predetermined development specifications. Rather than being used directly as the final instrument, this preliminary version underwent content validation by expert reviewers to ensure the appropriateness of its content, construction, and language before being administered to students. Following revisions based on the validators' recommendations, the instrument proceeded to empirical testing as part of the psychometric evaluation process. Consequently, this development stage not only produced an assessment instrument tailored to the characteristics of Akidah Akhlak Phase E learning but also established the foundation for the subsequent evaluation of validity, reliability, and item characteristics discussed in the following section.

Psychometric evaluation of the formative assessment instrument for higher order thinking skills

Psychometric evaluation was conducted to ensure that the developed *Higher Order Thinking Skills* (HOTS) formative assessment instrument met the requirements of a valid, reliable, and appropriate measurement tool for use in Akidah Akhlak Phase E learning. The evaluation encompassed three major aspects—content validity, empirical validity, and reliability—before proceeding to item characteristic analysis using the ITEMAN software package. These procedures were carried out sequentially to provide evidence that the instrument was not only conceptually appropriate but also demonstrated satisfactory performance based on empirical data. Consequently, the results of the psychometric evaluation served as the basis for determining the instrument's suitability before its implementation as a formative assessment for measuring students' higher-order thinking skills.

The first stage focused on evaluating content validity through expert judgment by two validators with expertise in educational assessment and Akidah Akhlak. Each item was assessed in terms of its alignment with the learning indicators, the HOTS construct, subject-matter content, and language appropriateness. The level of agreement between the validators was analyzed using Aiken's V coefficient to determine the adequacy of the content representation for each item. In addition to producing quantitative evidence, the validation process generated recommendations regarding item wording, stimulus clarity, and answer alternatives, which were subsequently used to refine the instrument before field testing.

Table 2. Content validity of the developed instrument

Category	Aiken's V Range	Number of Items
Very Valid	≥ 0.80	33
Valid	0.70–0.79	2
Fair	0.60–0.69	0
Poor	< 0.60	0
Total		35

Source: Developed by the Author (2026).

As presented in Table 2, the instrument achieved a mean Aiken's V coefficient of 0.846, with values ranging from 0.733 to 1.000. Of the 35 developed items, 33 were classified as *very valid*, while the remaining two were categorized as *valid*, and no item obtained a coefficient below 0.70. These findings indicate that all items adequately represented the HOTS indicators aligned with the learning objectives of Akidah Akhlak Phase E, making them suitable for empirical testing after minor revisions based on the validators' recommendations. Having established satisfactory content validity, the

subsequent stage examined whether each item also demonstrated acceptable measurement performance based on students' responses.

The next stage involved empirical validity testing using the Pearson Product-Moment correlation between individual item scores and the total test score. This analysis aimed to determine the extent to which each item represented the same construct measured by the overall instrument, ensuring that only items satisfying the statistical criteria were retained. The instrument was administered to 35 Phase E students, and the obtained correlation coefficients were compared with the critical r -value at the 5% significance level. This stage was essential because it provided empirical evidence that the quality of the instrument was supported not only by expert judgment but also by students' actual response data.

Table 3. Empirical validity test results

Criteria	Results
Number of Items Tested	35
Range of r -values	0.322–0.747
r -table ($\alpha = 0.05$; $N = 35$)	0.334
Valid Items	34
Invalid Item	1 (B21)
Decision	One item was removed, resulting in a final instrument of 34 items.

Source: Developed by the Author (2026).

The results presented in Table 3 indicate that 34 of the 35 items met the empirical validity criterion, whereas one item (B21) was eliminated because its correlation coefficient was below the critical r -value. With an acceptance rate of 97.14%, the instrument demonstrated a strong capacity to represent the HOTS construct based on empirical evidence. The removal of a single item was considered part of the refinement process to ensure that only items with optimal measurement quality were retained in the final instrument. Following the elimination of items that failed to meet the validity criterion, the instrument's quality was further evaluated through reliability testing to examine the internal consistency among the retained items.

Table 4. Instrument reliability test

Criteria	Results
Reliability Method	Cronbach's Alpha
Number of Items	34
Cronbach's Alpha	0.929
Interpretation	Very High
Decision	Reliable and suitable for use

Source: Developed by the Author (2026).

As shown in Table 4, the Cronbach's Alpha coefficient of 0.929 indicates that the instrument possesses a very high level of internal consistency and is capable of producing stable scores when administered to groups of students with comparable characteristics. This high reliability coefficient suggests that the items retained after empirical validity testing consistently measure the HOTS construct. Accordingly, the instrument satisfied the three primary indicators of psychometric quality: content validity, empirical validity, and reliability. Having fulfilled these requirements, the evaluation proceeded to an analysis of item characteristics using ITEMAN to obtain additional information regarding item difficulty, discrimination, and distractor effectiveness.

The characteristics of each test item were subsequently analyzed using the ITEMAN software package to provide additional psychometric evidence for the instrument. This analysis focused on three principal indicators in Classical Test Theory: the *difficulty index*, *discrimination index*, and *distractor effectiveness*. These indicators were employed to evaluate the quality of each item after the instrument had satisfied the validity and reliability requirements. Unlike the previous analyses, which assessed the quality of the instrument as a whole, the ITEMAN analysis provided detailed information about the performance of individual items in measuring students' competencies. Consequently, this analysis served as the basis for confirming that each item was not only conceptually appropriate but also demonstrated satisfactory measurement quality when implemented in formative assessment.

Table 5. Summary of item characteristics based on ITEMAN analysis

Analysis	Category	Number of Items	Percentage
Difficulty Index	Difficult	0	0.00%
	Moderate	21	61.76%
	Easy	13	38.24%
Discrimination Index	Very Good	30	88.24%
	Good	2	5.88%
	Fair	1	2.94%
	Poor	1	2.94%
Distractor Effectiveness	Effective	34	100.00%
	Ineffective	0	0.00%

Source: Developed by the Author (2026).

As presented in Table 5, most items demonstrated a *moderate* level of difficulty, accounting for 21 items (61.76%), whereas 13 items (38.24%) were classified as *easy*, and no items fell into the *difficult* category. This distribution indicates that the instrument possesses an appropriate level of difficulty, making it suitable for formative assessment intended to monitor students' learning progress. Regarding item discrimination, 30 items (88.24%) were categorized as *very good*, two as *good*, one as *fair*, and only one as *poor*. These findings demonstrate that the majority of the items effectively distinguished between students with high and low levels of ability, thereby contributing substantially to the accuracy of the measurement results.

In addition to exhibiting appropriate levels of difficulty and discrimination, the analysis also revealed that all distractors functioned effectively. The ITEMAN output identified no *flagged items*, indicating that every distractor successfully attracted students who had not yet mastered the material without reducing the likelihood that knowledgeable students would select the correct answer. Effective distractors constitute a critical component of well-designed multiple-choice tests because they enhance measurement precision while minimizing the possibility of students answering correctly by guessing. These findings further strengthen the technical quality of the instrument and demonstrate that the development of the answer alternatives was conducted systematically in accordance with established principles of educational measurement.

Overall, the psychometric evaluation demonstrates that the developed HOTS formative assessment instrument satisfies the essential requirements of a high-quality educational measurement instrument. Strong content validity confirms that the items accurately represent the intended construct, empirical validity indicates that nearly all items consistently measure HOTS competencies, and the very high reliability

coefficient demonstrates the stability of the measurement results. Furthermore, the ITEMAN analysis reinforces these findings by revealing an appropriate distribution of item difficulty, discrimination indices predominantly classified as *very good*, and fully effective distractors across all items. Collectively, these results indicate that the instrument possesses adequate psychometric quality and is suitable for use as a formative assessment instrument for measuring students' higher-order thinking skills in Akidah Akhlak Phase E.

Implications of instrument development for HOTS assessment in Akidah Akhlak

Although numerous studies have successfully developed *Higher Order Thinking Skills* (HOTS) assessment instruments, most have focused primarily on reporting validity and reliability without explaining how psychometric quality contributes to classroom assessment practices (Abadiga et al., 2025; Arafah et al., 2021; Zhou et al., 2024). Consequently, assessment instruments are often regarded merely as research products rather than educational tools capable of improving instructional quality. A similar condition is evident in Akidah Akhlak, where assessments predominantly emphasize conceptual knowledge rather than students' abilities to analyze, evaluate, and make decisions grounded in Islamic values. Therefore, this study not only develops an instrument that satisfies psychometric requirements but also proposes a conceptual framework illustrating the relationship among instrument development, psychometric evaluation, and its implementation in formative assessment, as summarized in Figure 1.

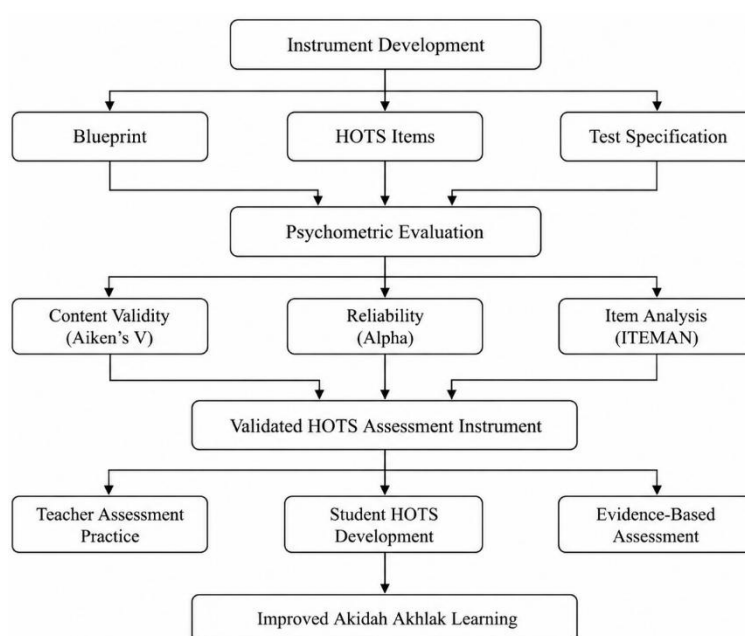


Figure 1. Conceptual framework of HOTS instrument development and its implications for formative assessment

Source: Developed by the Author (2026).

Figure 1 illustrates the interrelationship among instrument development, psychometric evaluation, and the implementation of formative assessment. The framework demonstrates that a high-quality assessment instrument is not established through a single evaluation stage but rather through a continuous process involving blueprint development, HOTS item construction, test specification, expert validation, and psychometric evaluation. This integrated process resulted in an instrument supported by strong evidence of validity and reliability, as indicated by a mean Aiken's V coefficient of 0.846, empirical validity for 34 of 35 items, a Cronbach's Alpha coefficient of 0.929, and fully effective distractors across all items. Accordingly, the framework illustrates that instrument quality emerges from the integration of systematic design procedures and comprehensive empirical verification.

Instrument Development constitutes the foundation for producing a high-quality HOTS assessment instrument because decisions made during the initial stages directly influence measurement quality throughout subsequent stages. The development of the blueprint, HOTS-based test items, and test specifications ensured that every item was explicitly aligned with the intended learning outcomes and competency indicators. This finding is consistent with the study by Connell et al. (2018), which argued that instrument quality is determined more by the rigor of the development process than by the complexity of statistical analyses conducted afterward. It also supports the findings of Hauer et al. (2023), who emphasized that systematically developed blueprints enhance construct representation while minimizing measurement error. Therefore, the high levels of content validity and reliability obtained in this study were not incidental but rather the logical outcome of a structured instrument development process.

Psychometric Evaluation serves as the critical link between instrument development and its practical implementation in educational settings. This stage ensures that each item is not only conceptually appropriate but also capable of producing accurate and consistent measurements based on empirical evidence. In the present study, psychometric quality was demonstrated by strong content validity (Aiken's V = 0.846), an empirical validity rate of 97.14% with 34 of 35 items meeting the required criteria, and very high reliability (Cronbach's Alpha = 0.929). These findings reinforce the conclusions of Albarqouni et al. (2018), who emphasized that educational instruments should be supported by evidence of validity and reliability before being implemented in instructional practice. Unlike Gupta (2023), whose evaluation focused primarily on validity and reliability, this study extends the evaluation by incorporating ITEMAN analysis, thereby providing a more comprehensive assessment of individual item quality.

The *Validated HOTS Assessment Instrument* represents the primary outcome that connects psychometric quality with classroom assessment practices. An instrument

that satisfies established validity and reliability standards provides teachers with a stronger foundation for implementing formative assessment because the resulting information accurately reflects students' higher-order thinking skills. This finding is consistent with Reynders et al. (2020), who reported that validated assessment instruments improve the accuracy of interpreting assessment results and consequently enhance the quality of feedback provided to students. However, this study extends the work of Clark and Watson (2019) by demonstrating that instrument quality depends not only on validity and reliability but also on sound item characteristics. This conclusion is supported by the predominance of moderately difficult items (61.76%), very good discrimination indices (88.24%), and 100% effective distractors, all of which contribute to more effective formative assessment.

The *Implementation in Formative Assessment of Akidah Akhlak Learning* demonstrates that an instrument meeting psychometric standards extends beyond being merely a research product and has direct implications for classroom practice. Teachers can employ the instrument to monitor students' HOTS development, provide more objective feedback, and design instructional follow-up activities based on assessment evidence. At the same time, students are encouraged to develop analytical, evaluative, and decision-making skills through contextualized assessment tasks that reflect the characteristics of Akidah Akhlak learning. Consequently, the principal contribution of this study lies not only in developing a valid and reliable assessment instrument but also in proposing a psychometric-based instrument development model that can serve as a reference for future HOTS assessment development across different subjects and educational levels.

Conclusion

This study successfully developed and evaluated a formative assessment instrument for *Higher Order Thinking Skills* (HOTS) in Akidah Akhlak Phase E through a systematic development process and comprehensive psychometric evaluation. The findings demonstrate that the instrument possesses excellent psychometric quality, as evidenced by a mean Aiken's V coefficient of 0.846, 34 of the 35 items (97.14%) satisfying the empirical validity criteria, a Cronbach's Alpha coefficient of 0.929, and item characteristics dominated by moderate difficulty, very good discrimination, and fully functioning distractors. These findings indicate that instrument quality is determined not only by statistical testing but also by a rigorous development process involving blueprint design, HOTS-based item construction aligned with the Learning Outcomes (*Capaian Pembelajaran* [CP]) and Learning Objectives (*Tujuan Pembelajaran* [TP]), expert validation, and refinement based on empirical evidence. Accordingly, this study demonstrates that integrating systematic instrument development with psychometric evaluation can produce an accurate formative assessment instrument for measuring students' higher-order thinking skills.

Beyond producing a valid and reliable assessment instrument, this study contributes a conceptual instrument development model that integrates instrument development, psychometric evaluation, and the implementation of formative assessment in Akidah Akhlak learning. The proposed model is expected to serve as a practical reference for teachers in designing evidence-based HOTS assessments and as a methodological framework for researchers developing educational instruments that meet psychometric standards. Nevertheless, this study is limited by the relatively small sample size and the implementation of the field test within a single school context, which may restrict the generalizability of the findings. Therefore, future studies are recommended to involve larger and more diverse samples, employ modern measurement approaches such as the *Rasch Model* or *Item Response Theory* (IRT), and extend instrument development to different learning materials and educational levels to enhance the applicability and generalizability of the proposed development model.

References

- Abadiga, Z. A., Tola, K. K., & Tulu, A. H. (2025). Development and validation of self-report questionnaire to assess higher-order thinking skills in secondary school mathematics. *School Science and Mathematics*. <https://doi.org/10.1111/ssm.18400>
- 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/10.1177/0013164485451012>
- Al-Htaybat, K., von Alberti-Alhtaybat, L., & Alhatabat, Z. (2018). Educating digital natives for the future: Accounting educators' evaluation of the accounting curriculum. *Accounting Education*, 27(4), 333–357. <https://doi.org/10.1080/09639284.2018.1437758>
- Albarqouni, L., Hoffmann, T., & Glasziou, P. (2018). Evidence-based practice educational intervention studies: A systematic review of what is taught and how it is measured. *BMC Medical Education*, 18(1), 177. <https://doi.org/10.1186/s12909-018-1284-1>
- Arafah, K., Amin, B. D., Sari, S. S., & Hakim, A. (2021). The development of higher order-thinking skills (HOTS) instrument assessment in physics study. *Journal of Physics: Conference Series*, 1899(1), 012140. <https://doi.org/10.1088/1742-6596/1899/1/012140>
- Arlianty, W. N., Febriana, B. W., Diniaty, A., & Fauzi'ah, L. (2018). Student profile in completing questions based on cognitive level of Bloom's taxonomy by Anderson and Krathwohl. *2nd International Conference on Chemistry, Chemical Process and Engineering (IC3PE)*, 020063. <https://doi.org/10.1063/1.5065023>

- Astin, A. W., & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Bloomsbury Publishing PLC.
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- Connell, J., Carlton, J., Grundy, A., Taylor Buck, E., Keetharuth, A. D., Ricketts, T., Barkham, M., Robotham, D., Rose, D., & Brazier, J. (2018). The importance of content and face validity in instrument development: Lessons learnt from service users when developing the Recovering Quality of Life measure (ReQoL). *Quality of Life Research*, 27(7), 1893–1902. <https://doi.org/10.1007/s11136-018-1847-y>
- Friatma, A., & Anhar, A. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series*, 1387(1), 012063. <https://doi.org/10.1088/1742-6596/1387/1/012063>
- Gärtner, F. R., Bomhof-Roordink, H., Smith, I. P., Scholl, I., Stiggelbout, A. M., & Pieterse, A. H. (2018). The quality of instruments to assess the process of shared decision making: A systematic review. *PLOS ONE*, 13(2), e0191747. <https://doi.org/10.1371/journal.pone.0191747>
- Gupta, K. (2023). Validity and reliability of students' assessment: Case for recognition as a unified concept of valid reliability. *International Journal of Applied & Basic Medical Research*, 13(3), 129–132. https://doi.org/10.4103/ijabmr.ijabmr_382_23
- Hauer, K. E., Park, Y. S., Bullock, J. L., & Tekian, A. (2023). "My assessments are biased!" Measurement and sociocultural approaches to achieve fairness in assessment in medical education. *Academic Medicine*, 98, S16–S27. <https://doi.org/10.1097/ACM.0000000000005245>
- Isnaeni, W., Rudyatmi, E., Ridlo, S., Ingesti, S., & Adiani, L. R. (2021). Improving students' communication skills and critical thinking ability with ICT-oriented problem-based learning and the assessment instruments with HOTS criteria on the immune system material. *Journal of Physics: Conference Series*, 1918(5), 052048. <https://doi.org/10.1088/1742-6596/1918/5/052048>
- Landers, R. N., Armstrong, M. B., Collmus, A. B., Mujcic, S., & Blaik, J. (2022). Theory-driven game-based assessment of general cognitive ability: Design theory, measurement, prediction of performance, and test fairness. *Journal of Applied Psychology*, 107(10), 1655–1677. <https://doi.org/10.1037/apl0000954>
- Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: The role of students' motivation. *Assessment &*

- Evaluation in Higher Education*, 46(2), 236–255.
<https://doi.org/10.1080/02602938.2020.1765228>
- Merle, M. H. (1998). Methodological studies: Instrument development. In *Advanced design in nursing research* (pp. 235–286).
- Reynders, G., Lantz, J., Ruder, S. M., Stanford, C. L., & Cole, R. S. (2020). Rubrics to assess critical thinking and information processing in undergraduate STEM courses. *International Journal of STEM Education*, 7(1), 9.
<https://doi.org/10.1186/s40594-020-00208-5>
- Safrudin, A. M. (2025). Learning assessment as a strategic tool for improving educational quality. *Islamic Journal of Education*, 4(2), 84–112.
<https://doi.org/10.54801/k22g7k67>
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2018). How to design a useful test. In *Understanding medical education* (pp. 275–289). Wiley.
<https://doi.org/10.1002/9781119373780.ch20>
- Solihin, R. R., Susanto, T. T. D., Fauziyah, E. P., Yanti, N. V. I., & Ramadhania, A. P. (2024). The efforts of Indonesian government in increasing teacher quality based on PISA result in 2022: A literature review. *Perspektif Ilmu Pendidikan*, 38(1), 57–65.
<https://doi.org/10.21009/PIP.381.6>
- Sugiyono, S., Lastariwati, B., Budiastuti, E., & Yudianto, A. (2018). Development of authentic assessment instruments for saintifical learning in tourism vocational high schools. *Jurnal Pendidikan Teknologi dan Kejuruan*, 24(1), 52–61.
<https://doi.org/10.21831/jptk.v24i1.16670>
- Zhou, Y., Ning, Y., Chen, J., Zhang, W., & Wijaya, T. T. (2024). Development and validation of Mathematical Higher-Order Thinking Scale for high school students. *Psychology in the Schools*, 61(8), 3160–3192. <https://doi.org/10.1002/pits.23213>