Tafkir: Interdisciplinary Journal of Islamic Education
Vol 6, No. 3, September 2025, DOI: https://doi.org/10.31538/tijie.v6i3.1940
E-ISSN 2527-8177 pp. 756-767

Development of A Comprehensive Instrument for The Islamic Education Subject Examination

Muh Syahrul Sarea*[1], Andi Harpeni Dewantara[2], Sabbar Dahham Sabbar[3]
[1] Institut Agama Islam Negeri Bone, Indonesia, syahrulsarea@iain-bone.ac.id
[2] Institut Agama Islam Negeri Bone, Indonesia, penidewantara@gmail.com
[3] Petra Educational Institute, Jordan, Sabbar.daham2000@gmail.com

Keywords:
Instrument Development, Comprehensive Examination, Islamic Religious Education.

Abstract

This study aims to develop a comprehensive instrument for the Islamic Education course examination. Because the implementation of the Comprehensive Examination is still partial, flexible, and not yet supported by an instrument that has gone through a systematic quality testing process, it employs a research and development (R&D) approach, adopting a modified version of the Borg and Gall development model, which consists of seven stages: (1) research and information gathering, (2) planning, (3) development of the initial product, (4) limited trial, (5) revision of the initial product, (6) field testing, and (7) final product revision. Data were collected through interview guidelines and questionnaires to examine the theoretical and empirical characteristics of the test items. Data analysis was conducted using the R Studio application to determine item discrimination, difficulty level, and distractor effectiveness. The results indicate that the final product consists of 5 difficult items (25%), 12 moderate items (60%), and three easy items (15%). In terms of item discrimination, eight items (40%) were classified as very good, seven items (35%) as good, and five items (25%) as requiring revision. Regarding distractor effectiveness, 17 items (85%) were categorized as good, while three items (15%) were considered requiring revision. Based on the development of comprehensive instruments, the Islamic Education course shows good quality and is suitable for use.

Kata kunci:
Pengembangan Instrumen, Ujian Komprehensif, Pendidikan Agama Islam.

Abstrak

Penelitian ini bertujuan untuk mengembangkan instrumen komprehensif pada mata ujian Ilmu Pendidikan Islam. karena pelaksanaan Ujian Komprehensif masih bersifat parsial, fleksibel dan belum didukung oleh instrumen yang telah melalui proses uji mutu yang sistematis. Penelitian ini merupakan penelitian pengembangan dengan menggunakan model pengembangan Borg and Gall yang telah dimodifikasi menjadi 7 tahapan pengembangan: (1) penelitian dan pengumpulan informasi, (2) perencanaan, (3) pengembangan produk awal, (4) uji coba terbatas, (5) revisi produk awal, (6) uji coba lapangan, dan (7) revisi produk akhir. Pengumpulan data menggunakan pedoman wawancara dan angket untuk melihat karakteristik butir secara teoritis dan empiris. Analisis data menggunakan bantuan aplikasi program R studio untuk melihat daya beda, tingkat kesukaran dan efektivitas distraktor. Hasil penelitian menunjukkan bahwa produk akhit menunjukkan bahwa butir sukar sebanyak 5 atau 25%, butir sedang 12 atau 60% dan butir mudah 3 atau 15%, sedangkan daya beda butir, sangat baik 8 atau 40%, baik 7 atau 35% dan kurang baik 5 atau 255. berdasarkan efektivitas distraktor diperoleh butir baik sebanyak 17 atau 85% dan kurang baik sebanyak 3 atau 15%. berdasarkan pengembangan instrumen Komprehensif mata kuliah ilmu Pendidikan islam menunjukkan kualias yang baik serta layak untuk digunakan

Article history:
Received: 28-04-2025
Revised 13-06-2025
Accepted 04-07-2025

Corresponding Author:
Muh Syahrul Sarea
Institut Agama Islam Negeri Bone, Indonesia, syahrulsarea@iain-bone.ac.id

INTRODUCTION

The comprehensive examination has become one of the main components in the higher education system for assessing students' competencies (Banta & Schneider, 1986). As an essential prerequisite for obtaining an academic degree, this examination is designed to assess students' understanding of the various materials studied throughout their coursework (Tobin & Gebo, 2008). The comprehensive examination is typically administered after the completion of all face-to-face courses. This is due to its holistic nature, as it encompasses material from various courses that the students have undertaken (Adela & Ritonga, 2023; Ahid & Chamid, 2021; AL-Momani, 2024; Arif et al., 2024; Auliyah et al., 2011).

In academic settings, comprehensive examinations play a significant role in ensuring that students possess an adequate mastery of their field of study (Anderson et al., 1984). Nevertheless, the implementation of this examination is often faced with various challenges (Al-Shanawani, 2019; Davis et al., 2022; Faraji et al., 2022). At IAIN Bone, the implementation of the comprehensive examination remains partial, with each faculty exercising its own policies in administering the exam. Examiners are granted the authority to independently design and develop the examination instruments. This flexibility is intended to facilitate both lecturers and students in conducting the examination (Alp Christ et al., 2022; Rahmawati et al., 2024). However, this partial approach may also result in variations in the quality of the instruments used, making it essential to evaluate and ensure that these instruments comply with prevailing academic standards (Damis, 2018).

One of the important comprehensive exams in the Faculty of Tarbiyah is the Islamic Education Science examination. This exam is not only mandatory but also aims to assess students' understanding of the foundational principles and theories in Islamic education. In this context, it is crucial to ensure that the instruments used are of high quality. A quality instrument should accurately portray students' abilities, identify their strengths and weaknesses, and provide useful information for future learning improvement (Krisma & Fatih'Adna, 2023) Unfortunately, the instruments currently in use have never been rigorously analyzed or quality-assessed by examiners, leaving their academic soundness in question. There is no standard instrument for comprehensive tests in the field of religion, especially Islamic education. The comprehensive instruments that currently exist only come from the exact sciences as research (Suwarna, 2016) who developed a comprehensive Physics test. Meanwhile, the instrument must have good item quality before being used.

Item quality is a critical factor in determining the validity and reliability of test instruments (Downing & Haladyna, 1997). Analyzing item quality is an essential step to ensure instruments maintain high levels of validity and consistent reliability. Furthermore, the instruments must meet quality parameters such as difficulty level, discrimination power, and distractor effectiveness (Gierl et al., 2017). The items' difficulty levels should be proportionately distributed across easy, moderate, and difficult categories to reflect the students' overall capabilities (Kholis, 2017). Item discrimination is an important indicator for assessing the extent to which an instrument can differentiate between high- and low-performing students (Johari et al., 2012). Additionally, each item's distractors must function effectively to challenge students and guide their choices accurately (Gierl et al., 2017).

High-quality instruments positively impact assessment outcomes and boost student confidence (Bray et al., 2020). Conversely, poorly designed items can lead to inaccurate evaluations, undermining the purpose of the assessment and reducing the

overall effectiveness of the learning process (Adetoro & Okike, 2022; Aseery, 2024; Colthorpe et al., 2021; Friatma & Anhar, 2019). Therefore, it is important to periodically analyze the items used in comprehensive exams to evaluate and improve instrument quality in alignment with academic needs.

In view of the importance of quality instruments, this study aims to develop a suitable comprehensive examination instrument for Islamic Education Science. In addition, currently there is no comprehensive instrument for the Islamic Education science exam subject while all PTKI use comprehensive exams for this subject. The resulting instrument is expected to meet the required quality standards, serving not only as an effective evaluation tool but also as a valuable source of data for enhancing the educational quality within the Faculty of Tarbiyah.

RESEARCH METHOD

This research is categorized as developmental research, aiming to produce a comprehensive instrument for the Islamic Education Science examination in the form of multiple-choice items. Developmental research focuses on the processes of designing, creating, validating, and refining a product to be used effectively in a specific context. In this case, the goal is to develop a high-quality academic evaluation instrument.

The development model applied is the Borg & Gall model, widely known as a systematic approach for educational product development (Gustiani, 2019). The model comprises ten main stages, including needs analysis, initial design, expert validation, revision, field testing, further revision, and implementation. As suggested by (Adib, 2017), researchers can adapt relevant stages based on the objectives of the study. This research uses a modified version of the Borg & Gall model into seven stages: Research and Information Gathering, Planning, Initial Product Development, Limited Testing, Large-Scale Testing, and Final Product Revision.

The study population consists of seventh-semester students of the Faculty of Tarbiyah from six study programs: Islamic Religious Education, Arabic Language Education, Islamic Education Management, English Language Education, Early Childhood Islamic Education, and Islamic Elementary School Teacher Education. A random sampling technique was used to ensure generalizability within the context of the Faculty. Four expert validators participated, including two subject experts in Islamic Education Science, one language expert, and one measurement expert. These validators assessed the instrument's quality in terms of content, language, and alignment with evaluation goals.

Various instruments were employed to gather both theoretical and empirical evaluation data. A key instrument was a validation form given to the expert validators. They assessed the instrument theoretically, including its language, construction, and content. In addition, test items were administered in two phases—limited and large-scale trials—to evaluate student responses.

Qualitative data were also collected through interviews and questionnaires. Interviews were conducted with course instructors and students to explore their experiences during the comprehensive exam. Questionnaires were used to gather students' perceptions of the instrument and identify challenges faced during the exam. This mixed-method approach ensured comprehensive data collection—covering theoretical, empirical, and contextual aspects.

Data analysis was performed descriptively using a quantitative approach via R version 2023.9.0.463 (Team, 2023). Empirical analysis included difficulty level, item discrimination, and distractor effectiveness. The criteria used are presented below:

Table 1. Criteria for Item Difficulty and Discrimination

Difficulty Level		Discrimination Index	
P ≤ 0,3	Difficult	0,4 ≤ D ≤ 1,0	Very Good
0,3 < P ≤ 0,7	Moderate	0,3 ≤ D ≤ 0, 3,9	Good
P > 0,7	Easy	0,2 ≤ D ≤ 0, 2,9	Require Revision
		D < 0,2	Poor

An instrument that meets empirical criteria is deemed suitable for use in the comprehensive examination, while instruments that do not meet the criteria are revised or replaced. The final instrument is expected to provide an accurate assessment of students' competencies in the Islamic Education subject.

RESEARCH RESULTS AND DISCUSSION

Research Results

The result of this research development is a comprehensive instrument for the Islamic Education subject. The development model employed is a modified version of the Borg & Gall model, utilizing only 7 out of the original 10 stages of development.

Research and Information Gathering

The initial stage of this research involved the collection of data and information regarding the implementation of comprehensive examinations at the Faculty of Tarbiyah. The researcher conducted interviews with lecturers responsible for the examination subjects as well as students who had previously taken the comprehensive exam. These interviews aimed to gain an understanding of the implementation procedures, the challenges encountered, and the characteristics of the instruments used thus far.

Based on the interview findings, it was revealed that the implementation of the comprehensive exam is characterized by a degree of flexibility. This flexibility is intended to facilitate both students and lecturers, particularly in adjusting examination schedules. Additionally, several campuses adopt varying formats of the comprehensive exam in accordance with their respective institutional policies (Bray et al., 2020). However, this condition also has implications for the evaluation standards, which tend to vary due to each lecturer having the autonomy to develop their own examination instruments. The instruments commonly take the form of subjective tests aimed at assessing students' conceptual understanding, given that the scope of the comprehensive exam material is quite broad (Ponder et al., 2004). Unfortunately, there has never been an in-depth analysis of the quality of these instruments, resulting in the validity and reliability of the instruments remaining unverified.

Moreover, the interview results indicate that, given the extensive scope of the exam material, the multiple-choice test format is more effective than other formats. Multiple-choice questions enable researchers to incorporate a wide range of content and competencies into a single comprehensive assessment tool (Özdemir & Toker, 2025). This serves as the basis for the researchers' decision to develop an instrument in the form of multiple-choice questions for the Islamic Education subject.

Planning

In the planning stage, the researcher conducted a classification of the material coverage to be assessed. Topics relevant to the Islamic Education subject were grouped based on themes, sub-themes, and levels of complexity. This process resulted in a blueprint or framework that serves as a guide for developing the instrument grid.

The blueprint includes essential elements such as content coverage, competency indicators, and the cognitive levels intended to be measured (Juliani et al., 2025). For

example, the competency indicators are designed to assess students' knowledge, understanding, and analytical abilities concerning the principles of Islamic education. In addition, the type of instrument, method of implementation, and cognitive levels are formulated in detail for each test item.

The researcher also designed a plan for analyzing the quality of the instrument, which includes the measurement of validity, reliability, as well as empirical analysis of parameters such as Difficulty Level, Discrimination Index, and distractor effectiveness. This planning aims to ensure that the developed instrument is not only theoretically sound but also capable of providing accurate empirical evaluations.

Initial Product Development

At this stage, the researcher developed test items based on the previously designed blueprint. A total of 20 items were constructed for the Islamic Education subject test. Each item was carefully designed to ensure alignment with the specified competency indicators and cognitive levels.

The developed instrument was then validated by a team of experts consisting of two subject matter experts, one language expert, and one assessment expert. The validation aimed to evaluate the instrument in terms of language, construction, and content relevance. The results of the expert validation are summarized in the following table:

Table 2. Characteristics of Comprehensive Items in Terms of Language, Construction, and Content

Subject	Number of Item	Good Item		
		Language	Construction	Material
Islamic Education	20	18	20	20

The validation results indicate that, in terms of language, 18 items were deemed appropriate, while the remaining 2 items required revision. The revisions were made to address the use of ambiguous wording that could lead to misinterpretation among students. Additionally, the distractors in several items were considered ineffective due to their overly extreme nature, thus failing to serve their purpose as plausible alternatives. The researcher revised both aspects based on the validators' feedback.

Limited trial

A limited trial was conducted on 20 items of the comprehensive instrument for the Islamic Education subject. The trial involved a number of students from the Faculty of Tarbiyah, with a set duration of 50 minutes for completing the test. This process aimed to evaluate the initial quality of the instrument through an analysis of item characteristics.

Difficulty Level Analysis

The Difficulty Level was determined using a formula that measures the proportion of participants who answered each item correctly. The analysis yielded three main categories: difficult, moderate, and easy items. The results of the Difficulty Level analysis are presented in the following table:

Table 3. Initial Product Difficulty Level

Categories	Item Number	Frequency	Percetage
P ≤ 0,3 (difficult)	9,10,12,13,15,16,18,20	8	40%
0,3 < P ≤ 0,7 (moderate)	1,2,3,4,5,7,11,14,17	9	45%
P > 0,7 (easy)	6,8,19	3	15%

The analysis results indicate that 40% of the test items fall under the difficult category, 45% are classified as moderate, and 15% fall under the easy category. This distribution of difficulty levels is considered sufficiently proportional for the purpose of comprehensive evaluation. However, several difficult items require review to ensure that the difficulty level does not introduce bias in the measurement of ability.

Discrimination Index Analysis

The Discrimination Index of each test item was calculated using the point-biserial correlation coefficient (*r pbis*), which measures the ability of an item to distinguish between high- and low-ability participants. The results of the Discrimination Index analysis are presented in the following table:

Table 4. Initial Product Discrimination Index

Category	Item Number	Frequency	Percentage
0,4 ≤ D ≤ 1,0 (very good)	3, 4, 5, 6	4	20%
0,3 ≤ D ≤ 0, 3,9 (good)	7, 8, 17,18, 20	5	25%
0,2 ≤ D ≤ 0, 2,9 (require revision)	2, 9, 10, 11, 16, 19.	6	30%
D < 0,2 (poor)	1, 12, 13, 14, 15,	5	25%

The analysis results indicate that 20% of the test items have an excellent Discrimination Index, 25% fall into the good category, 30% require revision, and 25% are considered poor. The items categorized as poor require comprehensive improvement to enhance discriminant validity.

Distractor Effectiveness Analysis

The effectiveness of distractors, or decoy options, was evaluated to determine whether the answer choices other than the correct key function effectively as distractors. The results of the distractor effectiveness analysis are presented in the following table:

Table 5. Distractor Effectiveness of the Initial Product

Category	Item Number	Frequency	Percentage
Rspp response answer key > 5%	2, 3, 4, 5, 7, 10, 12, 13, 14, 16, 17, 18, 20	13	65%
Rspp response answer key > 5%	1, 6, 8, 9, 11, 15, 19	7	35%

A total of 65% of the test items contained effective distractors, while the remaining 35% were require revision and require revision. The ineffective distractors in seven test items were revised to improve the quality of the instrument.

Overall Statistics of the Initial Product

The following table summarizes the instrument statistics based on the limited trial analysis:

Table 6. Statistical Results of the Initial Product Analysis

Statistic Scale	
nItem	20.000
nPerson	20.000
alpha	0.627
scaleMean	8.950
scaleSD	3.086

Large-Scale Field Testing

The field test was conducted with 47 students from the Faculty of Tarbiyah, representing six study programs. The duration for completing the test items remained limited to 50 minutes, as in the small-scale trial. This process aimed to ensure the consistency of the analysis results on a larger scale.

At this stage of the large-scale field testing, an analysis of the quality of the test items was also carried out. The quality of the items was assessed based on their characteristics, which included parameters such as the Difficulty Level, Discrimination Index, and the functionality of the distractors (Wibawa, 2019). This is further emphasized by (Quaigrain & Arhin, 2017) who state that a good instrument must meet several criteria: an appropriate ratio of items across different levels of difficulty, the presence of ideal items based on the Discrimination Index, and the effective functioning of distractors.

Difficulty Level Analysis

The Difficulty Level in the large-scale trial is presented in the following table: Statistical Results of the Main Product Analysis

Table 7. Difficulty Level of the initial product

Category	Item Number	Frequency	Percentage
P ≤ 0,3 (difficult)	9,10,12,13,16,	5	25%
0,3 < P ≤ 0,7 (moderate)	1,2,3,4,5,7,11,14,15,17,18, 20	12	60%
P > 0,7 (easy)	6,8,19	3	15%

The distribution of the Difficulty Level indicates a more ideal proportion, with 25% of the items categorized as difficult, 60% as moderate, and 15% as easy.

Discrimination Index Analysis

The results of the Discrimination Index analysis on a large scale are presented in the following table:

Table 8. Discrimination Index of the initial product

Category	Item Number	Frequency	Percentage
0,4 ≤ D ≤ 1,0 (very good)	3, 4, 5, 6, 7, 8, 17, 18	8	40%
0,3 ≤ D ≤ 0, 3,9 (good)	2, 11, 13, 14, 15, 19, 20	7	35%
0,2 ≤ D ≤ 0, 2,9 (require revision)	1, 9, 10, 12, 16,	5	25%
D < 0,2 (poor)	-	0	0%

The table above shows that there are 8 items, or 40%, with a very good Discrimination Index; 7 items, or 35%, with a good index; and 5 items, or 25%, with a poor index. There

are no items classified as having a very poor Discrimination Index. In addition to the Difficulty Level and Discrimination Index, the effectiveness of the distractors based on the distribution of responses is presented in the table below.

Table 9. The Effectiveness of Final Product Distractors

Category	Item Number	Frequency	Percentage
Rspp response answer key > 5%	1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14, 15, 16, 17, 18, 20	17	85%
Rspp response answer key > 5%	8, 11, 19	3	15%

In multiple-choice test instruments, distractors must be analyzed to evaluate whether the alternative answer choices, apart from the key, function effectively as distractors (Wibawa, 2019). According to (Mahjabeen et al., 2017), a distractor is considered effective if it is selected by more than 5% of test-takers. The table above shows that there are 17 items, or 85%, with effective distractors, while 7 items, or 15%, have require revision distractors. This indicates that the distractor options in these 7 items require revision. The revision is carried out by modifying or replacing the ineffective alternative answers in each relevant test item.

Final Prodcut Revision

The product revision was carried out to obtain a final version of the comprehensive examination instrument that possesses sound characteristics. Improvements were made based on expert input concerning language, construct, and content of the developed instrument. The initial version of the instrument contained ambiguous language, and some distractors were theoretically inappropriate, as they exhibited mainstream differentiation from other options, thereby limiting their effectiveness as functional distractors.

The second revision followed a small-scale trial and analysis using the R Studio program, focusing on item difficulty level, discrimination index, and distractor functionality. The data revealed that five items had poor discrimination indices, and seven items contained distractors that did not function effectively. Revisions were made by reviewing the formulation of the questions and answer choices in the affected items. The third revision, conducted after a large-scale trial, required minimal adjustments, as the items had already achieved satisfactory characteristics and were deemed suitable for use.

DISCUSSION

The development of a comprehensive instrument for the subject of Islamic Education at IAIN Bone was carried out through a modification of the Borg & Gall model, which is widely recognized in educational development research (Assyauqi, 2020). The selection of seven out of ten stages in this model reflects a contextual adaptation to the research needs and available resources, as suggested by (Gustiani, 2019), who stated that the Borg & Gall model may be modified according to the characteristics of the study and time constraints.

Research and Information Gathering

The initial stage of development focused on identifying the needs and the actual conditions of comprehensive exam implementation at the Faculty of Tarbiyah, IAIN Bone. Based on interviews with lecturers and students, it was found that the examination

process is flexible, and the instruments used are generally in the form of subjective multiple-choice tests. While this flexibility is perceived to facilitate the examination process, the instruments employed have never undergone a quality analysis. These findings indicate a gap between current practices and academic evaluation standards, which should prioritize validity and reliability. (Krisma & Fatih'Adna, 2023) emphasize that the quality of the instruments is crucial to ensure that comprehensive examinations can objectively assess students' competencies.

Planning

In the planning stage, the researcher classified the scope of materials relevant to Islamic Education to be subsequently organized into a blueprint. This blueprint includes the content scope, indicators, types of instruments, implementation methods, and cognitive levels for each test item. This step represents a systematic effort to ensure that every component of the instrument accurately reflects the competencies to be assessed. The planning phase also encompasses strategies for analyzing validity and reliability, as well as the criteria for high-quality test items, such as difficulty level, discrimination index, and distractor effectiveness. (Downing & Haladyna, 1997) emphasize the importance of developing a rigorous test blueprint as the foundation for content validity in multiple-choice instruments.

Initial Product Development

Referring to the blueprint and test specifications, the researcher developed 20 comprehensive test items, which were subsequently validated by experts in terms of language, construct, and content. The validation results indicated that 18 items met the language criteria, while 2 items required revision due to ambiguous wording and overly extreme distractors. In terms of construct and content, all items were deemed appropriate. Revisions were made based on expert feedback to ensure that the instrument has optimal discriminative power and avoids misinterpretation by respondents. (DiBattista & Kurzawa, 2011) stated that the quality of distractors significantly influences the effectiveness of test items in distinguishing students with different levels of competence. Therefore, this initial product meets the theoretical standards of a valid assessment instrument.

Limited Trial

At this stage, the comprehensive instrument on Islamic Education, consisting of 20 items, was piloted with 20 students over a 50-minute testing period. The analysis was conducted using the R Studio program to determine the item characteristics, including difficulty level, discriminating power, and distractor effectiveness. The results indicated that 40% of the items were categorized as difficult, 45% as moderate, and 15% as easy. In terms of discriminating power, only 20% of the items fell into the very good category, while 25% were considered poor and require revision. The distractor effectiveness analysis revealed that 65% of the items had well-functioning distractors, while the rest needed improvement. The internal reliability coefficient of the instrument at this stage was 0.627, indicating a moderate level of consistency for an initial stage, as described by (Ratumanan & Laurens, 2011)

Large-Scale Trial

At this stage, the test was administered to 47 students from six different study programs. The time and format of implementation remained the same as in the limited trial. The analysis results indicated improvements in several key aspects of the instrument. The item difficulty index showed a more balanced distribution: 25% difficult, 60% moderate, and 15% easy. The item discrimination index demonstrated significant improvement, with 40% categorized as very good and 35% as good, while no

items were found to be in the poor category. The effectiveness of distractors also improved, with 85% of the items containing functioning distractors as defined by (Mahjabeen et al., 2017), who stated that a distractor is considered effective if selected by more than 5% of test takers. Although the Cronbach's Alpha value was only 0.454, it still falls within the category of acceptable reliability, according to (Taber, 2018) indicating that the instrument can be used with an internally consistent level acceptable for developmental purposes.

Final Product Revision

The final stage, namely the revision of the final product, was carried out based on input from experts and the results of trial analysis. Revisions included improvements in language, construct, and content of several items identified as ambiguous or containing distractors that did not function optimally. In the limited trial, five items showed low discrimination power, and seven items had non-functioning distractors. Improvements were made by re-evaluating the wording of the questions and answer options. After the large-scale trial, the instrument demonstrated strong characteristics, and the final revisions were therefore only minor. The final results indicate that the instrument meets the criteria for content validity, distractor effectiveness, and acceptable internal reliability, making it suitable for use as a competency assessment tool for students in the Islamic Education subject examination.

CONCLUSION

Based on the analysis results, the comprehensive instrument in the Islamic Education subject demonstrates good quality and meets the eligibility criteria for use in implementing comprehensive exams. This instrument is the first standard in the field of religious sciences to have been developed systematically and tested in the context of formal academic evaluation. This finding marks the importance of strengthening the quality of assessment in the field of Islamic sciences. As a follow-up, further research is expected to expand the scope of the question bank, particularly in other Islamic Education subjects, to support the variation and depth of material evaluation. In addition, the development of a digital-based examination system that is easy to access and use is a strategic step to support efficiency, accuracy, and transparency in the implementation of comprehensive exams in the future.

REFERENCES

Adela, N., & Ritonga, A. A. (2023). The Effectiveness of The Ta'lim Program in Strengthening Islamic Religious Education for Students. *Nazhruna: Jurnal Pendidikan Islam*, *6*(3), Article 3. https://doi.org/10.31538/nzh.v6i3.3696

Adetoro, 'Niran, & Okike, B. (2022). Assessing Undergraduates Social competence on Social Media in Nigeria. *Library Philosophy and Practice (e-Journal)*. https://digitalcommons.unl.edu/libphilprac/6788

Adib, H. S. (2017). Teknik pengembangan instrumen penelitian ilmiah di perguruan tinggi keagamaan islam. *Prosiding Seminar Nasional & Internasional*.

Ahid, N., & Chamid, N. (2021). Implementation of Indonesian National Qualification Framework Based Curriculum in Higher Islamic Education. *Jurnal Pendidikan Islam*, *7*(1), Article 1. https://doi.org/10.15575/jpi.v7i1.12425

AL-Momani, M. O. (2024). The Degree of Parents' Practice of The Good Role Model Style Included in Islamic Educational Thought from The Point of View of University Students. *At-Tadzkir: Islamic Education Journal*, *3*(2), Article 2. https://doi.org/10.59373/attadzkir.v3i2.68

Alp Christ, A., Capon-Sieber, V., Grob, U., & Praetorius, A.-K. (2022). Learning processes and their mediating role between teaching quality and student achievement: A systematic review. *Studies in Educational Evaluation*, *75*, 101209. https://doi.org/10.1016/j.stueduc.2022.101209

Al-Shanawani, H. M. (2019). Evaluation of Self-Learning Curriculum for Kindergarten Using Stufflebeam's CIPP Model. *SAGE Open*, *9*(1), 2158244018822380. https://doi.org/10.1177/2158244018822380

Anderson, W. P., Krauskopf, C. J., Rogers, M. E., Neal, G. W., Rogers, M. E., & Neal, G. W. (1984). Reasons for comprehensive examinations: A re-evaluation. *Teaching of Psychology*, *11*(2), 78–82.

Arif, M., Chapakiya, S., & Dewi, A. Y. (2024). Character Education in Indonesia Islamic Elementary Schools: A Systematic Literature Review (2014-2024). *J-PAI: Jurnal Pendidikan Agama Islam*, *11*(1). https://ejournal.uin-malang.ac.id/index.php/jpai/article/view/29301

Aseery, A. (2024). Enhancing learners' motivation and engagement in religious education classes at elementary levels. *British Journal of Religious Education*, *46*(1), 43–58. https://doi.org/10.1080/01416200.2023.2256487

Assyauqi, M. I. (2020). Model Pengembangan Borg and Gall. *Researchgate, No. December*.

Auliyah, R., Herawati, N., & Utami, A. D. (2011). BAGAIMANAKAH PENAFSIRAN UJIAN KOMPREHENSIF MENURUT CIVITAS AKADEMIK UNIVERSITAS TRUNOJOYO? *InFestasi*, *7*(1), 64–75.

Banta, T. W., & Schneider, J. A. (1986). *Using Locally Developed Comprehensive Exams for Majors to Assess and Improve Academic Program Quality*.

Bray, A., Byrne, P., & O'Kelly, M. (2020). A short instrument for measuring students' confidence with 'key skills'(sicks): Development, validation and initial results. *Thinking Skills and Creativity*, *37*, 100700.

Colthorpe, K., Gray, H., Ainscough, L., & Ernst, H. (2021). Drivers for authenticity: Student approaches and responses to an authentic assessment task. *Assessment & Evaluation in Higher Education*, *46*(7), 995–1007. https://doi.org/10.1080/02602938.2020.1845298

Damis, R. (2018). Efektivitas Ujian Komprehensif Dalam Meningkatkan Kompotensi Mahasiswa Prodi Ilmu AqidaH. *Aqidah-Ta: Jurnal Ilmu Aqidah*, *4*(1), 57–72.

Davis, A., Meloncelli, N., Hannigan, A., & Ward, W. (2022). Evaluation of a model of online, facilitated, peer group supervision for dietitians working in eating disorders. *Journal of Eating Disorders*, *10*(1), 93. https://doi.org/10.1186/s40337-022-00617-7

DiBattista, D., & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), 4.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, *10*(1), 61–82.

Faraji, O., Ezadpour, M., Dastjerdi, A. R., & Dolatzarei, E. (2022). Conceptual structure of balanced scorecard research: A co-word analysis. *Evaluation and Program Planning*, *94*, 102128. https://doi.org/10.1016/j.evalprogplan.2022.102128

Friatma, A., & Anhar, A. (2019). Analysis of validity, reliability, discrimination, difficulty and distraction effectiveness in learning assessment. *Journal of Physics: Conference Series*, *1387*(1), 12063.

Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, *87*(6), 1082–1116.

Gustiani, S. (2019). Research and development (R&D) method as a model design in educational research and its alternatives. *Holistics*, *11*(2).

Johari, J., Abd Wahab, D., Ramli, R., Saibani, N., Sahari, J., & Muhamad, N. (2012). Identifying student-focused intervention programmes through discrimination index. *Procedia-Social and Behavioral Sciences*, *60*, 135–141.

Juliani, A., Nurhayati, N., Pertama, F. P., Nurhasanah, N., Hidayat, A., Iklimah, S. E., Herawati, N., Agustiani, T., Darmawan, A., & Puri, J. A. (2025). *Asesmen Multiliterasi*. Indonesia Emas Group.

Kholis, R. A. N. (2017). Analisis Tingkat Kesulitan (difficulty level) soal pada buku sejarah kebudayaan Islam Kurikulum 2013. *Jurnal Pendidikan Agama Islam*, *14*(2), 305–315.

Krisma, D. A., & Fatih'Adna, S. (2023). Analisis butir soal ujian tengah semester mata kuliah probabilitas: Bagaimana kualitasnya? *PYTHAGORAS: JURNAL PROGRAM STUDI PENDIDIKAN MATEMATIKA*, *12*(1), 1–15.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, *13*(4), 310–315.

Özdemir, A. Z., & Toker, Z. (2025). Analysis of distractors in mathematics questions and their potential to lead misconceptions. *Thinking Skills and Creativity*, *56*, 101730.

Ponder, N., Beatty, S. E., & Foxx, W. (2004). Doctoral comprehensive exams in marketing: Current practices and emerging perspectives. *Journal of Marketing Education*, *26*(3), 226–235.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013.

Rahmawati, R., Sastrawan, B., Martin, A. Y., Roestamy, M., Purnamasari, I., Maruapey, M. H., Ramdhani, M. R., & Danil, M. (2024). Assessing the Implementation of Kampus Mengajar Policy in Islamic Educational Institutions. *Jurnal Pendidikan Islam*, *10*(2), Article 2. https://doi.org/10.15575/jpi.v10i2.38976

Ratumanan, T. G., & Laurens, T. (2011). Penilaian hasil belajar pada tingkat satuan pendidikan. *Surabaya: Unesa*.

Suwarna, I. P. (2016). *Pengembangan instrumen ujian komprehensif mahasiswa melalui computer based test pada program studi pendidikan fisika*.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*, 1273–1296.

Team, P. (2023). RStudio: Integrated development environment for R. Posit Software. *PBC. Http://Www. Posit. Co*.

Tobin, K., & Gebo, E. (2008). Assessing student learning and departmental effectiveness through an undergraduate comprehensive exam. *Criminal Justice Studies*, *21*(3), 223–238.

Wibawa, E. A. (2019). Karakteristik butir soal tes ujian akhir semester hukum bisnis. *Jurnal Pendidikan Akuntansi Indonesia*, *17*(1), 86–96.