El-Rusyd
https://ejournal.stitahlussunnah.ac.id/index.php/el-rusyd

Evaluating the Quality of Objective and Essay Tests in Cognitive Assessment

Syamsi[1], Rahmat Junaidi[1], Burhanuddin Hidayat[1], Rozi Syofyandi[1], Julhadi[1]
[1]Universitas Muhammadiyah Sumatera Barat, Indonesia
sis680792@gmail.com *

Article Information:
Received October 15, 2025
Revised November 28, 2025
Accepted December 19, 2025

Keywords: *Cognitive assessment, assessment instruments, learning evaluation*

Abstract

Cognitive evaluation instruments are essential tools for systematically and objectively measuring students' thinking abilities. This study analyzes two primary types of cognitive assessment instruments objective tests and essay tests by evaluating their quality based on established indicators of validity and reliability. The purpose of this research is to examine how effectively these instruments measure cognitive performance across different levels of thinking skills. This study employs a qualitative research method with a content analysis approach, which involves reviewing, selecting, and analyzing scholarly literature, including textbooks, empirical studies, and theoretical documents related to test construction, validity, and reliability. The content analysis reveals that the cognitive domain encompasses mental processes such as problem-solving, reasoning, and the integration of conceptual knowledge. Written cognitive test instruments generally consist of essay tests (restricted and extended response) and objective tests (true–false, multiple-choice, matching, and completion). The findings demonstrate that the overall quality of cognitive evaluation instruments is largely determined by their validity and reliability, which serve as critical benchmarks for ensuring the accuracy and consistency of cognitive measurement.

INTRODUCTION

Assessment in the learning process always begins with a fundamental question: what should be assessed? This question is closely related to the main components of the teaching learning process, namely objectives, content, methods and tools, as well as assessment (Babullah, 2022; Htay et al., 2025; Rahman et al., 2025). Learning objectives are formulations of the behaviors that students are expected to master after engaging in learning experiences. Learning content refers to a set of scientific knowledge derived from the curriculum and delivered to students to ensure that the intended objectives are achieved (Asril et al., 2023; Djihadah et al., 2023). Methods and tools represent the techniques used to attain these objectives, while assessment serves to determine the extent to which learning goals have been achieved and to indicate the success of both the learning process and outcomes (Aran, 2024; Engkizar et al., 2023; Jami & Muharam, 2022).

How to cite: Syamsi, S., Junaidi, R., Hidayat, B., Syofyandi, R., Julhadi, J. (2025). Evaluating the Quality of Objective and Essay Tests in Cognitive Assessment. *El-Rusyd, 10*(2), 140-146.

E-ISSN: 2580-0256
Published by: The Institute for Research and Community Service

One of the primary domains evaluated in education is the cognitive domain, which encompasses mental activities related to knowledge, comprehension, and thinking processes (Kassymova et al., 2025; Mutiaramses et al., 2025; Stoltz et al., 2024). To measure achievement in this domain, evaluative instruments are required to transform abstract cognitive concepts such as understanding or analytical abilityinto empirical data that can be measured. Consequently, the quality of the data obtained largely depends on the quality of the instruments used (Roberts, 2020).

In general, cognitive test instruments consist of two main types: objective tests and essay tests. Objective tests such as multiple-choice, true–false, and matching items are widely used in large-scale assessments due to their efficiency; they allow for quick scoring, broad content coverage, and a high level of objectivity because each item has a definite correct answer (Imam et al., 2022; Yulastri et al., 2018). However, these tests are often criticized for primarily measuring lower-order thinking skills such as remembering (C1) and understanding (C2), and for allowing students to guess answers.

Conversely, essay tests are considered more effective in measuring higher-order thinking skills (HOTS), such as analyzing (C4), evaluating (C5), and creating (C6) (Eriyanti et al., 2022; Syafril, 2021; Wakifah et al., 2023). These tests require students to construct, organize, and express their ideas independently. Their main weakness lies in the potential for scoring subjectivityfor example, due to handwriting neatness, language style, or the halo effect (Engkizar et al., 2024; Engkizar et al., 2025; Engkizar et al., 2025).

The comparison between these two types of tests leads to an essential question: Do the instruments used truly measure what they are intended to measure, and do they produce consistent scores? Two fundamental concepts in psychometrics, namely validity and reliability, are central to answering this question. Instruments whether objective or essay-based that lack validity will produce inaccurate information, while those lacking reliability will yield inconsistent scores. Ignoring these two aspects may disadvantage students through unfair assessments and can mislead conclusions in educational research due to the use of poor-quality instruments.

METHODS

This study employs a qualitative method with a content analysis approach to evaluate the quality of objective and essay test instruments in cognitive assessment. This approach was chosen because the study focuses on examining various literature sources, including instructional evaluation textbooks, national and international journal articles, as well as theoretical documents discussing test construction, validity, and reliability (Baker et al., 2020; Langputeh et al., 2023; Pohontsch, 2019; Roller, 2019). Data were collected through documentation techniques by identifying, selecting, and reviewing relevant scholarly sources. The data analysis was conducted systematically using content analysis, which includes identifying data, coding key concepts, categorizing them into themes such as the characteristics of objective tests, essay tests, and indicators of validity and reliability, followed by interpreting the findings to understand patterns and relationships among these concepts. Through this approach, the study is able to conclude the quality of assessment instruments based on literature findings in an objective, in-depth, and structured manner (Muthatahirin et al., 2025; Pellegrini et al., 2021).

RESULT AND DISCUSSION

The Nature of Evaluation, Measurement, and Cognitive Testing

Evaluation in education is a planned, systematic, and continuous process used to determine the value and meaning of an object based on specific criteria as the

basis for decision-making. Although the terms assessment and evaluation are often used interchangeably, they differ in scope. Assessment focuses on a single component, such as student learning outcomes, whereas evaluation encompasses broader aspects of the educational system, including curriculum and instructional processes (Popham, 2017). This understanding aligns with Brookhart's (2018) view that evaluation functions as a mechanism for quality control in education as well as a tool to ensure that learning processes align with intended objectives.

According to Lessinger, evaluation is carried out by comparing learning objectives with students' actual achievements. The extent to which learning objectives are achieved is reflected in the fulfillment of competency standards established by the school (Popham, 2017). In the context of measurement, the term *test* originates from *testum*, meaning an instrument used to distinguish the value of an object. A test is a standardized instrument used to objectively measure learners' abilities or psychological conditions. Haladyna and Rodriguez (2013) similarly emphasize that tests are a central component of learning evaluation.

In educational evaluation, tests yield quantitative data on student abilities through responses to a series of tasks (Brookhart, 2010). A good test must meet the principles of utility, legality, feasibility, and measurement accuracy (Stiggins, 2014). Nitko and Brookhart (2014) stress that high-quality tests must satisfy both technical and substantive standards. International findings such as those by Rodriguez (2003) further show that the choice of test format significantly influences the accuracy of measuring cognitive ability. Numerous contemporary studies also highlight the importance of validity and reliability in assessment instruments. Haladyna and Rodriguez (2013) as well as Tavakol and Dennick (2011) affirm that test quality is strongly influenced by construct accuracy and consistency of measurement results.

In the cognitive domain, students' abilities are classified into six levels based on the revised Bloom's Taxonomy, ranging from remembering to creating (Brookhart, 2010). A good cognitive instrument must represent these levels comprehensively. Academic debates have emerged regarding the effectiveness of objective and essay tests in measuring thinking skills. Item analysis studies show that the quality of cognitive instruments is heavily influenced by item characteristics and construct alignment (Haladyna & Rodriguez, 2013). Similar findings are reported by McMillan (2013), who asserts that construct alignment is a fundamental component of classroom evaluation.

Objective tests require students to identify the correct answer from multiple options, whereas essay tests demand the construction of responses, involving more complex cognitive processes (Nitko & Brookhart, 2014). These findings are strengthened by Gulikers et al. (2004), who conclude that essay tests are more effective in assessing analysis and higher-order thinking. International studies such as Rodriguez (2003) also reveal that essay tests provide a more authentic representation of reasoning ability.

Characteristics, Strengths, and Weaknesses of Objective Tests

Objective tests include various formats, with multiple-choice items being the most commonly used. Each item consists of a stem, answer options, and distractors designed to measure ability accurately (Burton, 2005). The major strengths of objective tests include scoring objectivity, broad content coverage, and ease of item analysis. Brookhart (2010) notes that multiple-choice items can measure higher-order thinking when designed contextually.

However, objective tests also have weaknesses, such as the possibility of guessing and the difficulty of developing effective distractors. Downing (2005) emphasizes that guessing can reduce test reliability. Other studies also show that teacher-made tests often fail to meet item-writing standards, thereby reducing measurement accuracy (Stiggins, 2014).

Characteristics, Strengths, and Weaknesses of Essay Tests

Essay tests require students to construct their responses. Limited-response essays restrict the scope of answers, while extended-response essays provide students with full freedom to express their ideas (Nitko & Brookhart, 2014). Gulikers et al. (2004) affirm that essay tests are the most effective instruments for assessing higher-order thinking skills. This is consistent with Jonsson and Svingby (2007), who reveal that constructed-response instruments are more sensitive in measuring analytical and creative abilities.

The weaknesses of essay tests include scoring subjectivity, lengthy scoring time, and limited content coverage. Biases such as the halo effect often occur when scoring is conducted without rubrics. Therefore, analytic rubrics are essential, as emphasized by Brookhart (2018), who states that rubrics significantly enhance inter-rater consistency.

Pillars of Validity in Cognitive Testing

Validity determines whether an instrument truly measures the intended construct (Haladyna & Rodriguez, 2013). Content validity is judged through the representativeness of materials in the test blueprint and expert review (Nitko & Brookhart, 2014). In educational research, content validity has been shown to influence the appropriateness of items in measuring learning outcomes (Stiggins, 2014).

Construct validity ensures that the test reflects specific theoretical concepts. For objective tests, construct validity is often assessed through item–total correlations or factor analysis. Haladyna and Rodriguez (2013) highlight the importance of construct analysis in determining item quality. For essay tests, evidence of construct validity is obtained through theoretical justification and response-pattern analysis (Jonsson & Svingby, 2007).

Criterion validity examines the relationship between test scores and external measures, such as academic success indicators (Brookhart, 2010).

Pillars of Reliability in Cognitive Testing

Reliability refers to the consistency of measurement results. Objective tests with dichotomous scoring are generally evaluated using Kuder–Richardson-20 or split-half methods. Cortina (1993) states that coefficient alpha and Kuder–Richardson-20 are appropriate indicators of reliability because they consider item-level variance. Widana (as cited in Brookhart, 2010) emphasizes that reliability must be established before administering tests on a large scale.

In contrast, essay tests are evaluated using inter-rater reliability or Cronbach's alpha. Jonsson and Svingby (2007) as well as Brookhart (2018) demonstrate that the use of analytic rubrics significantly improves reliability. International findings such as Rodriguez (2003) also emphasize the importance of scoring consistency in evaluating constructed responses.

CONCLUSION

The findings of this study demonstrate that cognitive evaluation instruments fundamentally differ in their response processes and scoring mechanisms, thereby influencing the types of cognitive skills they can validly measure. Objective tests, characterized by their structured response format and high scoring objectivity, are effective for assessing a broad range of lower-order cognitive abilities efficiently. In contrast, essay tests provide deeper insights into learners' analytical, evaluative, and creative thinking due to their open-response format, although they require greater scoring time and carry higher risks of subjectivity. The analysis further confirms that validity particularly content and construct validity remains the principal foundation for ensuring the appropriateness of both objective and essay tests, while criterion validity supports their predictive strength. Reliability also emerges as a non-

negotiable requirement: objective tests demand internal consistency measures such as Kuder–Richardson-20, whereas essay tests rely heavily on inter-rater reliability supported by well-developed analytic rubrics. Collectively, these findings reinforce that the quality of cognitive assessment instruments depends on the alignment between test format, measurement objectives, and psychometric rigor, and that balanced use of both objective and essay tests is essential to capture the full spectrum of students' cognitive performance.

REFERENCES

Aran, D. (2024). Helping Future Schoolteachers Discover and Teach Soil: An Example of Project-Based Learning. *Spanish Journal of Soil Science*, *14*. https://doi.org/10.3389/sjss.2024.12280

Asril, Z., Syafril, S., Engkizar, E., & Arifin, Z. (2023). Advancing Educational Practices: Implementation and Impact of Virtual Reality in Islamic Religious Education. *Jurnal Pendidikan Islam*, *9*(2), 199–210. https://doi.org/10.15575/jpi.v9i2.20567

Babullah, R. (2022). Teori Perkembangan Kognitif Jean Piaget dan Penerapannya dalam Pembelajaran. *Epistemic: Jurnal Ilmiah Pendidikan*, *1*(2), 131–152. https://doi.org/10.70287/epistemic.v1i2.10

Baker, H. K., Kumar, S., & Pandey, N. (2020). A bibliometric analysis of managerial finance: a retrospective. *Managerial Finance*, *46*(11), 1495–1517. https://doi.org/10.1108/MF-06-2019-0277

Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. ASCD.

Brookhart, S. M. (2013). *Grading and group work: How do I assess individual learning?* ASCD.

Brookhart, S. M. (2018). *How to create and use rubrics for formative assessment and grading*. ASCD.

Burton, R. F. (2005). Multiple-choice and true/false tests: Myths and misconceptions. *Assessment & Evaluation in Higher Education, 30*(1), 65–72. https://doi.org/10.1080/026029304200324390

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104. https://doi.org/10.1037/0021-9010.78.1.98

Djihadah, N., Wasliman, I., Mulyanto, A., & Fatkhullah, F. K. (2023). Literary Teaching Based on Information and Communication Technology (ICT): An Inquiry Approach. *Theory and Practice in Language Studies*, *13*(6), 1556–1563. https://doi.org/10.17507/tpls.1306.25

Downing, S. M. (2005). The effects of guessing on multiple-choice reliability. *Applied Measurement in Education, 18*(2), 115–127. https://doi.org/10.1207/s15324818ame1802_2

Engkizar, E., Jaafar, A., Sarianto, D., Ayad, N., Rahman, A., Febriani, A., Oktavia, G., Guspita, R., & Rahman, I. (2024). Analysis of Quran Education Problems in Majority Muslim Countries. *International Journal of Islamic Studies Higher Education*, *3*(1), 65–80. https://doi.org/https://doi.org/10.24036/insight.v3i1.209

Engkizar, E, Jaafar, A., Masuwd, M. A., Rahman, I., Datres, D., Taufan, M., Akmal, F., Dasrizal, D., Oktavia, G., Yusrial, Y., & Febriani, A. (2025). Challenges and Steps in Living Quran and Hadith Research: An Introduction. *International Journal of Multidisciplinary Research of Higher Education (IJMURHICA*, *8*(3), 426–435. https://doi.org/10.24036/ijmurhica.v8i3.396

Engkizar, Engkizar, Jaafar, A., Taufan, M., Rahman, I., Oktavia, G., & Guspita, R. (2023). Quran Teacher: Future Profession or Devotion to the Ummah? *International Journal of Multidisciplinary Research of Higher Education (IJMURHICA)*, *6*(4), 196–210. https://doi.org/https://doi.org/10.24036/ijmurhica.v6i4.321

Engkizar, Engkizar, Muslim, H., Mulyadi, I., & Putra, Y. A. (2025). Ten Criteria for an Ideal Teacher to Memorize the Quran. *Journal of Theory and Research Memorization Quran*, *1*(1), 26–39. https://joqer.intischolar.id/index.php/joqer

Eriyanti, R. W., Cholily, Y. M., & Masduki, M. (2022). Meningkatkan Kreativitas Guru dalam Inovasi Pembelajaran Berbasis HOTS untuk Mengembangkan Berpikir Ktitis dan Kreatif Siswa. *To Maega : Jurnal Pengabdian Masyarakat*, *5*(3), 416. https://doi.org/10.35914/tomaega.v5i3.1176

Gulikers, J., Bastiaens, T., & Kirschner, P. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–86. https://doi.org/10.1007/BF02504676

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (3rd ed.). Routledge. https://doi.org/10.4324/9780203850381

Htay, S. S., Po, E. T. H., & Kaewkanlaya, P. (2025). Building Student Character through Worship in Elementary Schools. *Muaddib: Journal of Islamic Teaching and Learning*, *1*(2), 55–63. https://doi.org/https://muaddib.intischolar.id/index.php/muaddib/article/view/11

Imam, H., Hikmawati, Kosim, & Taufik, M. (2022). Pengaruh Model Pembelajaran Kooperatif Tipe Numbered Heads Together (NHT) Terhadap Hasil Belajar Siswa Kelas X SMAN 1 Sanggar Tahun Pelajaran 2021/2022. *Jurnal Pendidikan Fisika Dan Teknologi*, *8*(SpecialIssue), 58–66. https://doi.org/10.29303/jpft.v8iSpecialIssue.3715

Jami, D. Z., & Muharam, A. (2022). Strategy for Improving the Quality of Islamic Religious Education Study Programs with Total Quality Management. *Nidhomul Haq : Jurnal Manajemen Pendidikan Islam*, *7*(2), 267–283. https://doi.org/10.31538/ndh.v7i2.2096

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Kassymova, G. K., Talgatov, Y. K., Arpentieva, M. R., Abishev, A. R., & Menshikov, P. V. (2025). Artificial Intelligence in the Development of the Theory and Practices of Self-Directed Learning. *Multidisciplinary Journal of Thought and Research*, *1*(3), 66–79. https://mujoter.intischolar.id/index.php/mujoter/article/view/19

Langputeh, S., Andika, S., Ulfah, O., & Agusti, F. A. (2023). A Content Analysis: Values of Islamic Marriage in the Movie of Ayat-Ayat Cinta. *International Journal of Multidisciplinary Research of Higher Education*, *6*(3), 106–114. https://doi.org/10.24036/ijmurhica.v6i3.142

McMillan, J. H. (2013). Why we need research on classroom assessment. *Journal of Educational Measurement, 50*(4), 498–503. https://doi.org/10.1111/jedm.12044

Muthatahirin, M., Hanjit, C., Aminudin, W. S. A. B. W., & Nasir, A. A. B. A. (2025). Exploring Activities of International Dormitory Students to Advance Social Intelligence. *Journal of International Affairs and Students Mobility*, *1*(1), 17–28. https://doi.org/https://jiasmy.intischolar.id/index.php/jiasmy/article/view/2

Mutiaramses, M., Alkhaira, S., Zuryanty, Z., & Kharisna, F. (2025). Seven Motivations for Students Choosing to Major in Elementary School Teacher Education in Higher Education. *Multidisciplinary Journal of Thought and Research*, *1*(2), 23–37. https://mujoter.intischolar.id/index.php/mujoter/article/view/14

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (7th ed.). Pearson.

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective Programs in Elementary Mathematics: A Meta-Analysis. *AERA Open*, *7*. https://doi.org/10.1177/2332858420986211

Pohontsch, N. J. (2019). Qualitative Content Analysis. *Rehabilitation (Germany)*, *58*(6), 413–418. https://doi.org/10.1055/a-0801-5465

Popham, W. J. (2017). *Classroom assessment: What teachers need to know*. Pearson.

Rahman, F. A., Ulwi, K., & Aminudin, W. S. A. B. W. (2025). The Role of Islamic Education in Realizing in Sustainable Development Goals (SDGs 3). *Muaddib: Journal of Islamic Teaching and Learning*, *1*(2), 12–23. https://doi.org/https://muaddib.intischolar.id/index.php/muaddib/article/view/7

Roberts, R. E. (2020). Qualitative interview questions: Guidance for novice researchers. *Qualitative Report*, *25*(9), 3185–3203. https://doi.org/10.46743/2160-3715/2020.4640

Rodriguez, M. C. (2003). Constructed-response vs. multiple-choice tests: A review of the literature. *Practical Assessment, Research & Evaluation, 8*(1), 1–13. https://doi.org/10.7275/4m7k-0f52

Roller, M. R. (2019). A quality approach to qualitative content analysis: Similarities and differences compared to other qualitative methods. *Forum Qualitative Sozialforschung*, *20*(3). https://doi.org/10.17169/fqs-20.3.3385

Stiggins, R. J. (2014). *Revolutionize assessment: Empower students, inspire learning*. Corwin Press.

Stoltz, T., Weger, U., & da Veiga, M. (2024). Consciousness and education: contributions by Piaget, Vygotsky and Steiner. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1411415

Syafril, S. (2021). Learning Content and Process for Academically Talented Students. *Asian Social Science and Humanities Research Journal (ASHREJ)*, *3*(1), 73–81. https://doi.org/10.37698/ashrej.v3i1.64

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Wakifah, W., Fatimah, F., & Sulistiawati, M. (2023). Optimization of Higher-Order Thinking Skills (HOTS) in Islamic Education towards the Era of Society 5.0. *Didaktika : Jurnal Kependidikan*, *17*(2), 55–63. https://doi.org/10.30863/didaktika.v17i2.5750

Yulastri, A., Hidayat, H., Ganefri, G., Edya, F., & Islami, S. (2018). Learning outcomes with the application of product based entrepreneurship module in vocational higher education. *Jurnal Pendidikan Vokasi*, *8*(2), 120. https://doi.org/10.21831/jpv.v8i2.15310