International Journal of Education, Vocational and Social Science
https://e-journal.citakonsulindo.or.id/index.php/IJEVSS
E-ISSN : 2963 4954
IJEVSS

Development of Summative Assessment Instruments for Science in Elementary Schools Bontotiro District

Nurhapidah[1]
[1]Universitas Negeri Makassar, Indonesia
Baso Intang Sappaile[2]
[2]Universitas Negeri Makassar, Indonesia
Mansyur[3]
[3]Universitas Negeri Makassar, Indonesia

Corresponding Author: nurhapidah84@gmail.com ; baso.sappaile@unm.ac.id; mansyurunm@gmail.com

Abstract

This study aims to develop a valid and reliable summative assessment instrument for science learning in elementary schools in Bontotiro District. The research employed a development method involving systematic stages, including needs analysis, curriculum review, formulation of indicators, construction of test blueprints, item writing, expert validation, revision, pilot testing, item analysis, and finalization of the instrument. The results indicate that the developed instrument meets content validity criteria and is categorized as valid to highly valid, particularly in terms of alignment with the curriculum, conceptual accuracy, clarity of items, quality of answer keys, and appropriate language use, although several aspects such as distractor effectiveness and readability still require improvement. Furthermore, the item analysis shows that most items have moderate difficulty levels, adequate discrimination power, and functioning distractors, with only a few items needing revision or elimination. The reliability coefficient of the instrument is in the high category, indicating good internal consistency and stability in measuring student learning outcomes. Therefore, the instrument is considered feasible and reliable for use as a summative assessment tool in science learning at elementary schools in Bontotiro District.

Keywords: Development; Summative Assessment Instruments; Science in Elementary Schools; Bontotiro District

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

INTRODUCTION

The introduction of this study is grounded in the growing demands of 21st-century education, which emphasize the development of higher-order thinking skills, including critical thinking, creativity, communication, and collaboration. These competencies are not only essential requirements of modern curricula but also necessary skills for students to navigate the complexities of an increasingly dynamic global environment. In this context, the learning process must be designed to be meaningful, contextual, and oriented toward deep understanding. One of the key components that ensures the achievement of these learning objectives is assessment. Assessment plays a crucial role not only in measuring student learning outcomes but also in providing valuable information about students' progress and guiding instructional improvement. As highlighted in educational theory, assessment encompasses three main functions: assessment of learning, assessment for learning, and assessment as learning, all of which underscore the importance of designing valid, reliable, and comprehensive assessment instruments (Straub et al., 2017).

In science learning at the elementary school level, assessment holds a particularly important role in evaluating students' mastery of concepts, scientific process skills, and scientific attitudes. The current curriculum framework emphasizes that assessment should align with learning outcomes and support the development of students' scientific reasoning, problem-solving abilities, and reflective thinking. Therefore, assessment instruments in science education must go beyond measuring factual knowledge and instead capture conceptual understanding and the ability to apply knowledge in real-life contexts . However, preliminary observations conducted in several elementary schools in Bontotiro District reveal that the implementation of science summative assessment still faces significant challenges. Many teachers rely on self-developed instruments that have not undergone proper validation or reliability testing. As a result, the quality of test items is often questionable in terms of content accuracy, construction, and language clarity, which may lead to biased measurement outcomes. Furthermore, most of the questions tend to focus on lower-order cognitive skills, such as remembering and understanding, while higher-order thinking and scientific reasoning skills are rarely assessed (Alt, 2018).

Interviews with teachers further indicate that there is a lack of systematic guidance in developing summative assessment instruments aligned with learning outcomes. Teachers often adopt questions from various sources without ensuring their relevance to instructional objectives. In some cases, assessment instruments are constructed without prior analysis of alignment between test items and learning indicators. Additionally, the absence of item review processes

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

makes it difficult to accurately identify variations in students' abilities. This issue is reinforced by document analysis showing that students' score distributions tend to be homogeneous, with many students achieving high scores but struggling to explain concepts or apply them in practical situations. Such findings suggest that existing assessment instruments fail to measure students' scientific understanding comprehensively(Earle, 2014) .

From a psychometric perspective, the quality of an assessment is highly dependent on the quality of the instrument used. Instruments that do not meet validity and reliability standards can produce inaccurate and biased information. Classical Test Theory provides a comprehensive framework for evaluating instrument quality through analyses of theoretical validity, internal validity, difficulty index, discrimination power, distractor effectiveness, and reliability coefficients. These analyses enable educators to determine whether test items function effectively and whether the instrument as a whole can be trusted as a measurement tool . Nevertheless, findings from the preliminary study indicate that such psychometric analyses are rarely conducted by teachers in Bontotiro District. The lack of training and reference materials in quantitative analysis contributes to the limited ability of teachers to evaluate the quality of their assessment instruments, resulting in items that are either too easy, too difficult, or incapable of distinguishing between students of different ability levels(Erduran & Wooding, 2021) .

Moreover, the development of assessment instruments must consider the structure of learning outcomes. In grade V science for the first semester, learning outcomes consist of four main indicators that require adequate representation in test items. Ideally, a well-developed instrument should include a balanced number of items for each indicator, supported by a detailed blueprint to ensure alignment between content, cognitive level, and learning objectives. The development process should involve systematic stages, including blueprint construction, item writing, expert validation using Aiken's V, field testing with students, and item analysis based on Classical Test Theory principles (Gorin & Mislevy, 2013).

The situation in Bontotiro District highlights the urgent need for a standardized and well-developed summative assessment instrument in elementary science education. The absence of systematically developed instruments underscores the importance of conducting research that focuses on developing valid and reliable assessment tools. Such instruments are expected not only to provide accurate measurements of student learning outcomes but also to support teachers in improving instructional planning and implementation. Additionally, local geographical and socio-cultural conditions must be considered, as some schools face limitations in laboratory facilities and learning resources. In such contexts, well-designed written assessment instruments become

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

a practical and effective alternative for measuring learning outcomes objectively(Asiah, 2024) .

Based on these considerations, it is evident that the development of a valid, reliable, and psychometrically sound summative assessment instrument for science learning is a critical necessity. A high-quality instrument will provide accurate information about students' learning progress, support data-driven instructional decisions, and contribute to improving the quality of science education at the elementary level, particularly in Bontotiro District. Therefore, this study is conducted to bridge the gap between curriculum demands, theoretical principles of educational assessment, and actual assessment practices in schools, ultimately aiming to produce a standardized instrument that meets both theoretical and practical requirements.

The Role of Assessment in Science Learning at the Elementary Level

Assessment is a fundamental component of the learning process, particularly in science education at the elementary school level. It serves as a systematic process of collecting, interpreting, and utilizing information about students' learning progress. In modern educational paradigms, assessment is no longer viewed solely as a tool for measuring learning outcomes, but also as a mechanism to support and enhance the learning process itself. This perspective aligns with the concept that assessment functions in three major orientations: assessment of learning, assessment for learning, and assessment as learning. These three functions emphasize that assessment must not only evaluate final outcomes but also guide instruction and promote students' self-reflection(Kibble, 2017).

In the context of science learning, assessment plays a crucial role in measuring not only students' conceptual understanding but also their scientific process skills and attitudes. Science education is inherently multidimensional, encompassing knowledge, skills, and values. Therefore, assessment must be designed to capture this complexity. Effective science assessment should evaluate students' ability to observe, classify, interpret data, and draw conclusions, as well as their capacity to apply scientific concepts in real-life situations. This is particularly important in elementary education, where students are in the concrete operational stage of cognitive development and require meaningful and contextual learning experiences.

The implementation of assessment in science learning is also closely related to the current curriculum framework, which emphasizes competency-based education. The curriculum requires that assessment align with learning outcomes and support the development of higher-order thinking skills such as critical thinking, problem-solving, and scientific reasoning. Consequently, assessment instruments must be carefully constructed to reflect these competencies. However,

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

in practice, many teachers still rely on traditional forms of assessment that focus primarily on memorization and recall. This approach limits the ability of assessment to capture deeper levels of understanding and often fails to provide accurate information about students' true abilities(Daka et al., 2021).

Another important aspect of assessment in science education is its role in providing feedback. Assessment results should be used to inform both teachers and students about learning progress and areas that need improvement. For teachers, assessment data can guide instructional planning and help identify effective teaching strategies. For students, assessment provides an opportunity to reflect on their learning and develop metacognitive skills. Therefore, assessment must be designed not only as a summative tool but also as an integral part of the learning process.

Despite its importance, the quality of assessment in elementary science education often remains a concern. Many assessment practices do not meet the standards of validity and reliability, which are essential for producing accurate and meaningful results. Without proper validation and analysis, assessment instruments may fail to measure the intended competencies and may lead to biased or misleading conclusions. This highlights the need for systematic development and evaluation of assessment instruments to ensure their quality and effectiveness(Ani et al., 2025).

Development of Summative Assessment Instruments Based on Classical Test Theory

The development of assessment instruments is a critical process that determines the quality of measurement in educational evaluation. In the context of summative assessment, the goal is to measure students' learning outcomes at the end of a learning period accurately and objectively. To achieve this, assessment instruments must be developed through a systematic and rigorous process grounded in sound theoretical principles. One of the most widely used frameworks for this purpose is Classical Test Theory (CTT).

Classical Test Theory provides a foundation for understanding the relationship between observed scores, true scores, and measurement errors. According to this theory, the score obtained by a student is a combination of their true ability and random error. Therefore, the primary objective of instrument development is to minimize measurement error and ensure that the instrument accurately reflects students' true abilities. This requires careful attention to various psychometric properties, including validity, reliability, and item characteristics(Marzano, 2010).

The development process typically begins with the analysis of learning outcomes and the

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

construction of a test blueprint. The blueprint serves as a guide for ensuring that all relevant competencies and content areas are adequately represented in the assessment. It outlines the relationship between learning indicators, content domains, and cognitive levels, providing a structured framework for item development. This step is crucial for maintaining content validity and ensuring that the assessment aligns with instructional objectives.

Following the development of the blueprint, test items are constructed according to established guidelines for writing multiple-choice questions. Each item must be clear, concise, and free from ambiguity, with plausible distractors that effectively differentiate between students of different ability levels. The quality of item construction significantly influences the overall effectiveness of the assessment instrument(Earle, 2020).

Once the items are developed, they undergo expert validation to evaluate their content relevance, clarity, and appropriateness. This process often involves the use of Aiken's V coefficient to quantify the level of agreement among experts. Items that do not meet the required validity criteria are revised or eliminated. This step ensures that the instrument has strong theoretical validity before being tested empirically.

The next stage involves field testing the instrument with a representative sample of students. The data collected from this process are analyzed to evaluate the empirical characteristics of each item. Key indicators include item difficulty, discrimination index, and distractor effectiveness. Item difficulty measures the proportion of students who answer the item correctly, while discrimination index assesses the ability of the item to differentiate between high- and low-performing students. Distractor analysis examines whether incorrect answer options function effectively in attracting students who do not know the correct answer.

Reliability analysis is also conducted to determine the consistency of the instrument. In multiple-choice tests, reliability is commonly calculated using the Kuder-Richardson (KR-20) formula. A high reliability coefficient indicates that the instrument produces stable and consistent results. Together, these analyses provide a comprehensive evaluation of the instrument's quality(Harris et al., 2008).

Despite the availability of these methods, many teachers lack the knowledge and skills to conduct systematic instrument development and analysis. As a result, assessment instruments are often used without proper validation, leading to inaccurate measurement of student learning outcomes. This underscores the importance of research and development efforts aimed at producing standardized and psychometrically sound assessment instruments that can be used effectively in educational settings.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Characteristics of Effective Science Learning and Implications for Assessment Design

Science learning at the elementary school level is designed to help students understand natural phenomena through direct experience and inquiry-based activities. It encompasses three main dimensions: scientific knowledge, scientific processes, and scientific attitudes. These dimensions reflect the holistic nature of science education, which aims to develop not only cognitive understanding but also practical skills and positive dispositions toward learning.

One of the defining characteristics of science learning is its emphasis on the scientific method. Students are encouraged to observe, ask questions, conduct experiments, analyze data, and draw conclusions. This process helps develop critical thinking and problem-solving skills, which are essential for scientific literacy. Therefore, assessment in science education must be aligned with these learning processes and should evaluate students' ability to engage in scientific inquiry(Kerr et al., 2016).

Another important characteristic is the contextual nature of science learning. Concepts are often related to real-life situations and phenomena that students encounter in their daily lives. This approach enhances students' understanding and makes learning more meaningful. As a result, assessment instruments should incorporate contextual stimuli that reflect students' experiences and environments. Contextualized questions not only improve comprehension but also enable students to apply their knowledge in practical situations.

The cognitive development of elementary school students also has important implications for assessment design. At this stage, students are generally in the concrete operational phase, meaning they learn best through tangible and observable experiences. Therefore, assessment items should be designed to match their cognitive level, using clear language and familiar contexts. Complex or abstract questions may hinder students' ability to demonstrate their true understanding(Wilson & Toyama, 2018).

In addition to cognitive aspects, science learning also emphasizes the development of scientific attitudes such as curiosity, honesty, and objectivity. While these aspects are more difficult to measure through traditional tests, they should still be considered in the overall assessment strategy. Although this study focuses on multiple-choice tests, it is important to recognize that comprehensive assessment in science education should include a variety of methods to capture all dimensions of learning.

However, the implementation of effective science assessment often faces practical challenges, particularly in schools with limited resources. Constraints such as lack of laboratory facilities, limited teaching materials, and insufficient teacher training can hinder the use of

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

performance-based assessments. In such contexts, well-designed written assessments become a practical alternative for evaluating student learning. These assessments must be carefully constructed to ensure that they remain valid, reliable, and aligned with learning objectives(Castleberry et al., 2016).

In conclusion, the characteristics of science learning provide important guidance for the design of assessment instruments. Effective assessment must be aligned with learning objectives, reflect the nature of scientific inquiry, and consider students' cognitive and contextual conditions. By incorporating these principles, assessment instruments can provide a more accurate and comprehensive picture of students' learning outcomes and contribute to the improvement of science education at the elementary level.

DISCUSSION

The Reliability of the Instrument is at a High Level, so the Instrument is Feasible to be Used for Summative Assessment

The main results of this study indicate that the science summative assessment instrument has a KR-20 reliability coefficient of 0.721. This value indicates that the instrument is in the high category based on the criteria for interpreting reliability in Classical Test Theory. Thus, the instrument has demonstrated an adequate level of internal consistency in measuring the same construct. This confirms that the test items work relatively uniformly in representing the science learning outcomes of fifth-grade students.

This finding directly answers the problem formulation that focuses on the level of instrument reliability. The results obtained indicate that the reliability is at a level sufficiently large to support its use in summative assessment. Good internal consistency is an indicator that the instrument is able to provide stable results. Therefore, the instrument can be considered feasible for use in the context of learning evaluation(Ismail et al., 2022).

From the perspective of educational measurement, reliability is closely related to the stability and trustworthiness of scores. The research proposal has emphasized that KR-20 is an appropriate coefficient for multiple-choice tests with dichotomous scoring. In addition, a threshold value of ≥ 0.70 is considered the minimum standard to indicate acceptable internal consistency. Thus, the achievement of 0.721 indicates that the instrument has met and even exceeded these criteria.

The significance of this finding is very high because reliability is a primary prerequisite in educational assessment. Scores generated by a reliable instrument can be used as a basis for

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

making academic decisions more accurately. Without adequate reliability, assessment results have the potential to be influenced by measurement errors. Therefore, this finding reinforces that the instrument already has quality that can be scientifically justified(Gallardo, 2020).

The KR-20 Value of 0.721 Indicates Internal Consistency is Adequate, but Not Yet Maximum

Although the reliability value of 0.721 is classified as high, this achievement has not yet reached the very high category. This indicates that the instrument is already quite good, but still has room for improvement. In other words, the instrument has met the minimum standard, but has not yet reached the optimal level in terms of internal consistency. Therefore, the interpretation of this value needs to be carried out proportionally.

Psychometrically, values in this range indicate that most of the score variation reflects real differences in students' abilities. However, there is still the possibility of measurement errors that have not been completely minimized. This condition is reasonable in the development of educational instruments, especially at the initial implementation stage. Therefore, this result needs to be viewed as both a good achievement and an opportunity for improvement(de Vries et al., 2018).

Reliability is basically a continuous concept, not a binary category. An instrument with a reliability of 0.72 is certainly different in quality compared to an instrument that has reliability above 0.85. Although both can be categorized as reliable, the level of confidence in the scores produced remains different. This emphasizes the importance of continuous efforts to improve the quality of the instrument.

The urgency of this finding lies in its implications for further instrument development. Reliability values should not be seen as a static final result. Instead, these values must become the basis for carrying out revisions to items that are still weak. Thus, the instrument can continue to improve in quality through a continuous evaluation cycle(Black et al., 2010).

The Components of p, q, and Total Score Variance Support the Formation of Good Reliability

The analysis results show that the reliability of the instrument is influenced by several important components, namely the number of items, the distribution of p and q, and the variance of the total score. In this study, the instrument consists of 24 items with a total of 144 respondents. The value of Σpq of 5.380 and the total score variance of 17.415 indicate a fairly good distribution of responses. The combination of these factors contributes to achieving high reliability.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Within the framework of Classical Test Theory, reliability increases when there is sufficient variation in scores among students. In addition, items that have a moderate level of difficulty also tend to increase internal consistency. The results of the study indicate that most items are not in the extreme category. This allows the instrument to differentiate student abilities more effectively(Brownlie et al., 2024).

A sufficiently wide distribution of scores indicates that the instrument is able to capture variations in student abilities. This is important because reliability is highly dependent on such variation. If the scores are too homogeneous, then the ability of the instrument to differentiate participants will decrease. Therefore, a good score distribution is an important indicator in supporting reliability.

The practical implication of this finding is the importance of paying attention to the quality of item distribution in instrument construction. Teachers not only need to pay attention to the number of questions, but also the characteristics of each item. By understanding the relationship between p, q, and total score variance, improvements to the instrument can be carried out in a more directed manner. This shows that reliability is the result of the overall quality of test design(Dolin et al., 2017).

Reliability Strengthens the Finding that the Majority of Items Have Worked Quite Well

The finding of reliability of 0.721 also strengthens the results of the analysis. In that analysis, it is known that most items have adequate validity. In addition, many items are at a moderate level of difficulty and have good discrimination power. This condition directly contributes to the high reliability of the instrument.

Reliability cannot be separated from the quality of individual test items. An instrument consisting of low-quality items will be difficult to achieve high reliability. Conversely, if the majority of items have good characteristics, then overall reliability will increase. This shows a close relationship between item analysis and reliability.Analysis at the item level and test level complement each other in providing an overview of instrument quality. Thus, the interpretation of reliability becomes more comprehensive. This strengthens the validity of the conclusions produced.The urgency of this finding lies in the importance of conducting integrated analysis. Reliability is not sufficient to be understood only as a single number. Instead, it needs to be seen in the context of the quality of its constituent items. This approach provides a deeper understanding of instrument quality(Leuenberger, 2001).

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

The Reliability Finding is Very Urgent for the Bontotiro Context, because it Answers the Gap in Assessment Practices in the Field

In the context of Bontotiro District, this reliability finding has very high practical value. So far, teachers tend to construct questions without adequate psychometric analysis. This causes the quality of assessment to be unable to be evaluated objectively. Therefore, the results of this study provide a real contribution to assessment practices in the field.An instrument with a reliability of 0.721 indicates that the assessment developed is not only practical, but also has a strong scientific basis. This is important to improve the quality of assessment in elementary schools. Teachers can use this instrument with a higher level of confidence. Thus, assessment results become more meaningful(Edwards, 2017).

The aspect of fairness in assessment is also better guaranteed with a reliable instrument. The scores produced better reflect students' actual abilities. This reduces the possibility of errors in making academic decisions. Therefore, reliability has direct implications for the quality of learning. This finding also bridges the gap between theory and practice. The instrument developed not only meets academic standards, but is also relevant to field conditions. Thus, this study provides a significant contribution to improving the quality of basic education in the area(Looney, 2011).

Comparison of Findings with Previous Research: Similarities

The results of this study show conformity with various previous studies. The use of KR-20 as a reliability measurement tool is in line with the principles in Classical Test Theory. This is also consistent with the framework that has been explained in the research proposal. Thus, the methodological approach used has a strong theoretical basis.In addition, the interpretation of reliability as an indicator of score trustworthiness is also consistent with the views of Nitko & Brookhart and Kubiszyn & Borich. Reliability is considered a primary requirement so that test results can be interpreted validly. This shows similarity in conceptual foundations. Thus, this study strengthens existing theory.

This finding is also in line with other CTT-based instrument development studies. In those studies, reliability is considered the final result of the validation and item analysis process. The process carried out in this study follows these stages. Therefore, the results obtained can be said to be consistent with previous research.This similarity shows that the study has good external validity. The results obtained do not stand alone, but are supported by relevant literature. This strengthens confidence in the research findings. Thus, the contribution of the study becomes

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

increasingly significant(Jacobs et al., 2008).

Comparison of Findings with Previous Research: Differences

Although it has many similarities, this study also shows several important differences. One of the main differences lies in the context of instrument development. The instrument developed is specific to fifth-grade students in Bontotiro District. This provides strong contextual value to the research results.Another difference lies in the interpretation of the level of reliability. This study emphasizes that high reliability is sufficient for the context of classroom assessment. Not all instruments must achieve very high reliability to be usable. This approach is more realistic and contextual.

In addition, this study links reliability with detailed analysis of item characteristics. This approach provides a deeper understanding compared to studies that only report reliability values. Thus, the interpretation of results becomes more comprehensive. This is a particular advantage of this study(Alt, 2018).

These differences indicate the unique contribution of this study. The results obtained not only confirm existing theory, but also provide new perspectives. Thus, this study enriches studies in the field of educational evaluation. This makes it relevant both theoretically and practically.

Implications of Findings for the Development of Science Summative Assessment Instruments

The main implication of this finding is that the instrument developed is feasible to be used in summative assessment. A reliability of 0.721 indicates that the instrument is able to produce consistent scores. This provides a strong basis for its use in learning evaluation. Thus, the instrument can be adopted by teachers with a high level of confidence.

Practically, teachers in Bontotiro District can utilize this instrument as a reliable evaluation tool. This instrument is not only easy to use, but also has adequate psychometric quality. This improves the quality of the assessment process in the classroom. Thus, learning can be evaluated more objectively.However, the instrument still needs to be continuously refined. High reliability does not mean that it cannot be improved. Revisions to less optimal items are still necessary. This is important to maintain the quality of the instrument in the long term.Methodologically, this study emphasizes the importance of reliability analysis in instrument development. Without such analysis, the quality of scores cannot be ensured. Therefore, a data-based evaluation culture needs to be continuously developed. This finding provides an important contribution in encouraging more scientific assessment practices in elementary schools(Gorin & Mislevy, 2013).

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

CONCLUSION

The development process of the science summative assessment instrument for elementary schools in Bontotiro District was conducted systematically, beginning with needs analysis and curriculum review, followed by formulating indicators, designing the instrument blueprint, writing test items, expert validation, revision, pilot testing, item analysis, and final product revision supported by adequate documentation. The theoretical validation results indicated that the instrument met content feasibility criteria and was categorized as valid to highly valid, with strengths in curriculum alignment, conceptual accuracy, clarity of test items, answer key quality, and appropriate language use, although some aspects such as answer options, distractor effectiveness, readability for elementary students, and wording clarity still require improvement. Overall, the item characteristics showed relatively good quality, as most items fulfilled internal validity criteria, were dominated by moderate difficulty levels, demonstrated adequate discrimination power, and were supported by effective distractors, despite several items still needing revision due to being too easy, having low discrimination power, or ineffective distractors, with one item requiring elimination. Furthermore, the developed instrument achieved a high reliability category, indicating good internal consistency and dependability in measuring student learning outcomes, making it not only content-feasible but also sufficiently reliable as an assessment tool for science learning in elementary schools in Bontotiro District.

FUTURE DIRECTIONS

Future directions of this study should focus on strengthening the empirical quality and broader applicability of the developed science summative assessment instrument by conducting larger-scale field testing across diverse school contexts to enhance generalizability and external validity. Further refinement is needed through continuous item revision, particularly improving distractor effectiveness, readability, and alignment with higher-order thinking skills. In addition, integrating more advanced psychometric approaches such as Item Response Theory could provide deeper insights into item performance and student ability estimation. Future research should also explore the integration of digital-based assessment formats to support more efficient implementation and data analysis. Moreover, expanding the instrument to include performance-based and process-oriented assessments would allow a more comprehensive evaluation of students' scientific skills and attitudes. Finally, professional development programs for teachers on assessment design and analysis are essential to ensure sustainable use and ongoing improvement of high-quality assessment practices in elementary science education.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

REFERENCES

Alt, D. (2018). Teachers' Practices In Science Learning Environments And Their Use Of Formative And Summative Assessment Tasks. Learning Environments Research, 21(3), 387–406.

Ani, M., Obodo, A., & Enyi, N. (2025). The Practice Of Formative And Summative Assessments In Enhancing Basic Science Teaching And Learning. Journal Of Innovations In Educational Assessment, 7(2), 21–37.

Asiah, S. (2024). Development Of Diagnostic, Formative And Summative Assessment Instruments In The Pjbl Model To Stimulate Students' Critical And Creative Thinking Skills. Jurnal Penelitian Pendidikan Ipa, 10(10), 7486–7492.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity In Teachers' Summative Assessments. Assessment In Education: Principles, Policy & Practice, 17(2), 215–232.

Brownlie, N., Burke, K., & Van Der Laan, L. (2024). Quality Indicators Of Effective Teacher-Created Summative Assessment. Quality Assurance In Education, 32(1), 30–45.

Castleberry, A. N., Schneider, E. F., Carle, M. H., & Stowe, C. D. (2016). Development Of A Summative Examination With Subject Matter Expert Validation. American Journal Of Pharmaceutical Education, 80(2), 29.

Daka, H., Mulenga-Hagane, M. L., Mukalula-Kalumbi, M., & Lisulo, S. (2021). Making Summative Assessment Effective.

De Vries, A. H., Muijtjens, A. M. M., Van Genugten, H. G. J., Hendrikx, A. J. M., Koldewijn, E. L., Schout, B. M. A., Van Der Vleuten, C. P. M., Wagner, C., Tjiam, I. M., & Van Merriënboer, J. J. G. (2018). Development And Validation Of The Toco–Turbt Tool: A Summative Assessment Tool That Measures Surgical Competency In Transurethral Resection Of Bladder Tumour. Surgical Endoscopy, 32(12), 4923–4931.

Dolin, J., Black, P., Harlen, W., & Tiberghien, A. (2017). Exploring Relations Between Formative And Summative Assessment. In Transforming Assessment: Through An Interplay Between Practice, Research And Policy (Pp. 53–80). Springer.

Earle, S. (2014). Formative And Summative Assessment Of Science In English Primary Schools: Evidence From The Primary Science Quality Mark. Research In Science & Technological Education, 32(2), 216–228.

Earle, S. (2020). Balancing The Demands Of Validity And Reliability In Practice: Case Study Of A Changing System Of Primary Science Summative Assessment. London Review Of Education, 18(2), 221–235.

Edwards, F. (2017). The Development Of Summative Assessment Literacy: An Exploration Of The

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Experiences Of Beginner Secondary Science Teachers In New Zealand. The University Of Waikato.

Erduran, S., & Wooding, S. J. (2021). A Project Calibrate Approach To Summative Assessment Of Practical Science. School Science Review, 102(381), 71–77.

Gallardo, K. (2020). The Importance Of Assessment Literacy: Formative And Summative Assessment Instruments And Techniques. In Workgroups Eassessment: Planning, Implementing And Analysing Frameworks (Pp. 3–25). Springer.

Gorin, J. S., & Mislevy, R. J. (2013). Inherent Measurement Challenges In The Next Generation Science Standards For Both Formative And Summative Assessment. Invitational Research Symposium On Science Assessment.

Harris, K., Bauer, M., & Redman, M. (2008). Cognitive Based Developmental Models Used As A Link Between Formative And Summative Assessment. Proceedings Of The 34th International Association For Educational Assessment (Iaea) Annual Conference, Cambridge, England, September 7, 12.

Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative Vs. Summative Assessment: Impacts On Academic Motivation, Attitude Toward Learning, Test Anxiety, And Self-Regulation Skill. Language Testing In Asia, 12(1), 40.

Jacobs, C. L., Martin, S. N., & Otieno, T. C. (2008). A Science Lesson Plan Analysis Instrument For Formative And Summative Program Evaluation Of A Teacher Education Program. Science Education, 92(6), 1096–1126.

Kerr, S., Muller, D., Mckinon, W., & Mc Inerney, P. (2016). An Online Formative Assessment Tool To Prepare Students For Summative Assessment In Physiology. African Journal Of Health Professions Education, 8(1), 72–76.

Kibble, J. D. (2017). Best Practices In Summative Assessment. Advances In Physiology Education, 41(1), 110–119.

Leuenberger, T. (2001). Developing And Using Diagnostic And Summative Assessments To Determine Students' Conceptual Understanding In A Junior High School Earth Science Classroom. In Assessment In Science: A Guide To Professional Development And Classroom Practice (Pp. 197–209). Springer.

Looney, J. W. (2011). Integrating Formative And Summative Assessment: Progress Toward A Seamless System?

Marzano, R. J. (2010). Formative Versus Summative Assessments As Measures Of Student Learning. In Handbook Of Data-Based Decision Making In Education (Pp. 259–271).

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Routledge.

Straub, J., Marsh, R. A., & Whalen, D. J. (2017). Summative Assessment. In Small Spacecraft Development Project-Based Learning: Implementation And Assessment Of An Academic Program (Pp. 151–173). Springer.

Wilson, M., & Toyama, Y. (2018). Formative And Summative Assessments In Science And Literacy Integrated Curricula: A Suggested Alternative Approach. In Language, Literacy, And Learning In The Stem Disciplines (Pp. 231–260). Routledge.

Copyright at authors some right reserved this work is licensed under a Creative Commons Attribution-Share Alike 4.0 International License