 ISSN 2776-5229

Development of Semester Final Examination (UAS) Test Instruments Based on Higher Order Thinking Skills (HOTS) in the IPAS Subject for Fourth Grade Elementary School

Nisrayanti[1*], Baso Intang Sappaile[2], Mansyur[3]
[1]Universitas Negeri Makassar, Indonesia
[*1]nisrayanti1@gmail.com

ARTICLE INFO

Article history
Received April 3,2026
Revisied May 14, 2026
Accepted May 16,2026

Keywords: Test Instruments, HOTS, IPAS, End-of-Semester Exam, Elementary School.

ABSTRAK

This research aims to develop a Semester Final Examination (UAS) test instrument based on Higher Order Thinking Skills (HOTS) for the IPAS subject in the 4th grade of Elementary School. The background of this research is the still low ability of students' higher-order thinking, caused by the dominance of Lower Order Thinking Skills (LOTS) based questions in learning evaluation. Therefore, it is necessary to develop an assessment instrument that can measure students' critical thinking, analytical, and problem-solving abilities. This research uses the Research and Development (R&D) method with a development model that includes the stages of needs analysis, design, product development, expert validation, testing, and product revision. Validation was conducted by 2 validators consisting of subject matter experts and evaluation experts, while the instrument trial was carried out on 30 fourth-grade elementary school students. The validation results showed that the instrument received a very valid category with a feasibility percentage of 89%, while the reliability results indicated a reliability coefficient of 0.87 with a high category. The developed instrument consists of HOTS-based UAS questions on the IPAS material for fourth-grade elementary school students and is deemed suitable for use after undergoing the validation and revision process. This instrument is also deemed capable of measuring students' higher-order thinking skills more effectively compared to conventional instruments. The main contribution of this research is to produce a valid, reliable, and practical HOTS-based evaluation instrument to improve the quality of IPAS learning assessments in elementary schools.

1. INTRODUCTION

Higher Order Thinking Skills (HOTS) are abilities that are highly needed in the 21st century, where learners are required to analyze, evaluate, and create in solving various complex problems. According to the revised Bloom's Taxonomy, HOTS are at the cognitive levels C4 (analyzing), C5 (evaluating), and C6 (creating) (Tanujaya et al., 2017). HOTS-based learning and assessment encourage students to think critically, creatively, and reflectively about the knowledge they acquire (Brookhart, 2010; Conklin & Materials, 2012). A 10-year-old child is at the final stage of the concrete operational phase (7-12 years) before transitioning to the formal operational phase. At this stage, the child has developed the ability to think logically about concrete objects and is beginning to show strategic thinking skills. (Piaget, J, 1952). Based on the developmental characteristics of 10-year-old children, they already have the basic ability to think at a higher level (able to think analytically about concrete objects, able to evaluate information logically, starting to be able to think creatively in familiar contexts). Understanding this developmental theory has important implications in designing HOTS assessment instruments, such as using concrete contexts familiar to students, linking them to students' direct experiences, and also considering social-emotional development in question formulation. In the context of the Merdeka Curriculum, one of the important innovations implemented is the Natural and Social Sciences (IPAS) subject, which

integrates elements of both Natural Sciences and Social Sciences (Hattarina et al., 2022). The goal is for students to gain an integrated learning experience between natural and social phenomena, while also training their scientific, collaborative, and contextual thinking skills. However, the implementation of IPAS in elementary schools faces significant challenges in the preparation and execution of high-quality HOTS-based assessments. In many elementary schools, the assessments used are still dominated by Lower Order Thinking Skills (LOTS) questions, such as recalling or understanding facts. Teachers still struggle to design questions that measure analytical, evaluative, and creative abilities due to limited training and technical references. Just like in the island regions, facing limitations in access to training and educational technology. Syahrul et al. (2025) noted that teachers in this region do not yet have adequate competence in analyzing the quality of test items based on classical theory. As a result, the exam questions used are rarely tested for validity, reliability, or discrimination power, so the effectiveness of the instrument is not accurately known. The lack of continuous training also causes teachers to rely more on personal experience in creating questions. Research by Hulaipah, Syukri, and Indraswati (2023) revealed that fourth and fifth-grade elementary school teachers still face difficulties in understanding the concept of HOTS, formulating HOTS-based questions, linking questions to learning outcomes, and assessing the alignment with assessment standards. This condition indicates the need for the development of systematic, standardized HOTS test instruments that can be used as a reference for teachers. From the students' perspective, Lamiah (2025) found that fourth-grade students had difficulty completing HOTS IPAS questions, particularly in the aspects of analyzing, evaluating, and creating. Most students scored below the Minimum Completeness Criteria (KKM), especially when asked to connect concepts or create diagrams. This indicates that higher-order thinking skills have not developed optimally due to limited exposure to valid HOTS questions. These findings are supported by Rini and Rufi'i (2023), who conducted an analysis of the IPAS class IV SD question items in the chapter "Stories About My Region." The results indicate that the item validity is still low, reliability is moderate, the difficulty level tends to be easy, and the discrimination power is good. This means that most of the questions do not meet the criteria for quality instruments and need to be improved to align with HOTS learning objectives.

Research by Saputra, Asrin, and Novitasari (2024) also shows that although HOTS-based learning tools have been implemented in several elementary schools, their implementation has not been optimal due to constraints in facilities, teaching strategies, and time management. This emphasizes the importance of providing HOTS test instruments that can help teachers objectively assess students' higher-order thinking skills. Meanwhile, Raras, Siswanto, and Wijayanti (2024) proved that the use of Higher Order Thinking Skills (HOTS)-based Student Worksheets (LKPD) effectively improves the learning outcomes of fourth-grade elementary school students in IPAS. Similarly, Fitria, Wijaya, and Danial (2020) developed HOTS-based LKPD using the 4D model, which was proven to be valid, practical, and effective in training higher-order thinking skills. Both studies demonstrate the effectiveness of the HOTS approach, but their focus is on learning devices, not on assessment instruments. At the secondary education level, Sinaga and Yusuf (2023) developed a HOTS-based test instrument on acid-base material using the ADDIE model. The results are valid based on the Aiken's V index, but the study has not yet included practicality and construct validity tests. Thus, the research at the high school level can serve as a methodological reference, but it still leaves room for further development in more in-depth psychometric analysis. Although various studies on HOTS have been conducted, there are still several limitations that have not been addressed. Most previous research has focused more on the development of learning tools such as LKPD or HOTS modules, while the development of HOTS assessment instruments in elementary school IPAS subjects is still relatively limited. Research that develops HOTS test

instruments generally only emphasizes one aspect of testing, such as content validity or Rasch analysis, without integrating construct validity, reliability, item quality analysis, and instrument practicality comprehensively. In addition, some of the instruments that have been developed have not yet been adapted to the characteristics of elementary school students, especially in the context of island regions with limited educational resources and teacher training access. As a result, the available instruments are not yet fully capable of accurately, objectively, and contextually measuring students' higher-order thinking skills. This research differs from previous studies because it develops a HOTS-based UAS test instrument for the IPAS subject in the 4th grade of elementary school with a more comprehensive testing approach. This research not only tests content validity using Aiken's V but also examines construct validity thru Confirmatory Factor Analysis (CFA), internal reliability, item quality, and the practicality of the instrument thru responses from teachers and students. Furthermore, this research was developed based on the learning context in island regions, making the resulting instrument more contextual and relevant to the actual conditions of elementary schools. Thus, this research is expected to fill the gaps in previous studies while also producing a valid, reliable, practical, and applicable HOTS assessment instrument in the implementation of the Merdeka Curriculum.

2. METHODS

This research is a development study (Research and Development/R&D) aimed at producing a product in the form of an IPAS test instrument based on Higher Order Thinking Skills (HOTS) for fourth-grade elementary school students in the island region of Selayar. This development research approach is used because it focuses on the process of designing, testing, and validating instruments that are suitable for measuring students' higher-order thinking skills. The development model used refers to the 4D model proposed by Thiagarajan, Semmel, and Semmel (1974). This research is a Research and Development (R&D) study aimed at producing and developing a Higher Order Thinking Skills (HOTS)-based Semester Final Examination (UAS) test instrument for the fourth-grade IPAS subject in elementary school. The resulting product is expected to be used as an evaluation tool capable of measuring students' higher-order thinking skills more optimally compared to conventional instruments.

3. RESULTS AND DISCUSSION

The Process of Developing Semester Final Examination (UAS) Test Instruments Based on Higher Order Thinking Skills (HOTS) for the IPAS Subject in Grade IV Elementary Schools in Island Regions

The define stage in this research is aimed at identifying the need for the development of HOTS-based UAS IPAS test instruments at the elementary school level. In this stage, an analysis is conducted on the initial conditions of assessment implementation, student characteristics, and the relevance of the competencies to be measured in IPAS learning. This analysis process serves as the conceptual foundation for formulating the instrument design in the next stage. Thus, the define stage functions as an exploratory phase that emphasizes the urgency of instrument development while also determining the direction of the development design. The research results indicate that the development of HOTS-based UAS IPAS test instruments for fourth-grade elementary school students in island regions is a real and urgent need. The main findings in the early stages of the research show that assessment practices in schools are still dominated by the use of LOTS questions. This condition indicates that the assessments conducted do not fully support the development of higher-order thinking skills in students. These findings are significant because they affirm that the issues faced are not only due to the limitations of standard instruments but also the

gap between curriculum demands and the actual evaluation practices occurring in schools. In the research proposal, it is explained that IPAS learning in the Merdeka Curriculum is aimed at developing critical, creative, and reflective thinking skills in students. However, the needs analysis results indicate that teachers still tend to use questions oriented toward lower-order thinking skills. The discrepancy between the curriculum objectives and the assessment practices indicates an implementation challenge at the school level. Therefore, the development of HOTS-based test instruments is an important step to bridge the gap between curriculum demands and the assessment practices occurring in the classroom. Other findings indicate that teachers face difficulties in formulating HOTS indicators, designing contextual stimuli, and creating logical and well-functioning distractors. These difficulties indicate that the development of instruments cannot solely focus on providing a set of questions. The developed product needs to be designed comprehensively so that it can be operationally used by teachers. Thus, the resulting instrument must include a systematic guide from indicator mapping to the utilization of test results.

The urgency of developing comprehensive instruments is also related to the sustainability of product utilization in schools. If the developed product consists only of a collection of questions without accompanying development rationale and usage guidelines, teachers are likely to revert to simpler conventional assessment patterns. Therefore, the developed product needs to provide conceptual and practical support for teachers in implementing HOTS-based assessments. This approach is expected to encourage more sustainable changes in assessment practices. Findings regarding student characteristics also make an important contribution to the instrument development process. The analysis results show that fourth-grade students in the island region find it easier to understand stimuli that are visual, concise, and concrete. These characteristics indicate that the formulation of HOTS questions at the elementary school level needs to consider the cognitive development stage of the students. In other words, the application of HOTS to elementary school students cannot be done by simply copying the HOTS question models used at higher education levels. These findings emphasize that the development of HOTS questions for elementary school students must still consider the simplicity of language and clarity of context. The designed questions need to use stimuli that are easily understood by students, but still require a deep reasoning process. This shows that the difficulty level of the questions is not determined by the complexity of the language or the length of the reading. On the contrary, the quality of HOTS is more determined by the demands of analysis, evaluation, and decision-making built thru stimuli relevant to student development. If compared to previous research referenced in the proposal, there are several similarities in the orientation of product development. This research, like previous studies, places HOTS as the main objective in the development of learning devices or assessments. In addition, this research is also in line with Fitria, Wijaya, and Danial (2020) who used the 4D model and showed that HOTS-based products can meet the criteria of being valid, practical, and effective. Another similarity lies in the development approach carried out in stages, starting from needs analysis to product refinement. However, there are several fundamental differences between this study and previous research. The research by Fitria, Wijaya, and Danial focuses on the development of LKPD, whereas this study emphasizes the development of UAS test instruments. Thus, the product produced in this research is not a learning device, but rather an assessment tool that must meet certain psychometric standards. In addition, this research specifically places the geographical context of archipelagos as the basis for instrument design, while many previous studies have not given special attention to this aspect.

Another difference can be seen when this research is compared to the study by Sinaga and Yusuf (2023). Both studies are oriented toward the development of HOTS-based instruments, but they use different development models. The study by Sinaga and Yusuf uses the ADDIE model and

has not comprehensively examined the aspects of practicality or construct validity. On the other hand, this research uses the 4D model and was designed from the outset to integrate content validity testing, construct validity, reliability, item quality analysis, and instrument practicality into a systematic development process. Based on the research results and discussion, the development process of the UAS IPAS test instrument based on HOTS for fourth-grade elementary school students in the island region was carried out thru the 4D model, starting with the identification of real needs in the field. The define stage shows the dominance of LOTS question usage, the limitations of teachers in creating HOTS questions, the need for stimuli that match student characteristics, and the importance of integrating the local island context. The design stage then translates these needs into the preparation of blueprints, mapping cognitive levels C4–C6, multiple-choice test formats with four options, and the use of contextual stimuli. Next, the develop stage focuses the revision process on aspects of language, cognitive level accuracy, distractor quality, and local context suitability, while the disseminate stage indicates the product's readiness for limited use by teachers. Thus, the development process results in a strong foundation for the creation of contextual, systematic, and relevant UAS HOTS IPAS instruments that meet the learning needs in elementary schools in island regions.

Results of the Content Validity Test of the UAS HOTS IPAS Instrument for 4th Grade Elementary School

The content validity test of the UAS HOTS IPAS instrument for 4th-grade elementary school was conducted thru an assessment process by four validators with relevant expertise backgrounds. The four validators consisted of IPAS subject matter experts, educational evaluation or measurement experts, IPAS elementary school teachers working in island regions, and 4th-grade elementary school teachers. The assessment process was carried out using a 1–4 rating scale that encompasses several important aspects, namely material suitability, relevance to learning outcomes or indicators, accuracy of HOTS level, question construction, language use, contextual relevance, and clarity of instructions. Methodologically, this procedure aligns with the research design and the content validation sheet used, which places content validity as an expert assessment of the alignment of test items with competencies, learning indicators, HOTS principles, question construction, language, and context relevant to the characteristics of island regions. The research results show that the content validity of the UAS HOTS IPAS test instrument for 4th-grade elementary school is in the quite strong category with an average Aiken's V value of 0.756. These findings provide an answer to the question regarding the feasibility of the content of the developed instrument. The value indicates that, in general, the instrument has met the basic requirements of content validity, although not all items have reached a very high level of validity. Thus, this finding confirms that the instrument has a sufficient development foundation, but still requires improvements in several parts to ensure that all items truly meet the expected standards. The findings on content validity lead to two main implications in the instrument development process. First, these results indicate that the direction of instrument development has been on the right track according to the research design. Second, the revision process is still necessary to ensure that all items are fully consistent with the HOTS product specifications outlined in the research proposal. The proposal explicitly states that the final product must meet content validity standards as evidenced by the Aiken's V index. Therefore, items that have not reached the feasibility threshold must be revised before being retained as part of the final instrument. The finding that eight out of fifteen items are in the very valid category indicates that most of the instrument's content aligns with the scope of the 4th-grade IPAS material and the UAS assessment objectives. This condition also suggests that the item construction has adhered to the general principles of educational test construction. The appropriateness is important because content validity is initial evidence that the

instrument does not deviate from the domain of competence to be measured. In the research proposal, it is emphasized that content validity serves to ensure that each item truly represents the content and measurement objectives set. The presence of eight highly valid items indicates that the developed product has a sufficiently strong conceptual foundation. This reflects that most of the learning indicators have been accurately translated into the form of questions. In addition, the alignment between indicators, content, and assessment objectives has been systematically considered in the development process. Thus, these results show that some of the instruments have been able to adequately represent the measurement construct. However, the more important finding is related to the presence of six items that fall into the less valid category and one item in the fairly valid category. This condition indicates that several items still require refinement to fully meet the content validity standards. The main issue does not lie in the appropriateness of the subject matter, but rather in the accuracy of the expected cognitive demands. In other words, the content of the lesson presented in that item is correct, but the demand for higher-order thinking has not been fully realized. Some items still remain at a relatively low cognitive level compared to the HOTS instrument development targets. These questions tend to require the ability to remember or understand basic information, rather than the ability to analyze, evaluate, or create. This condition indicates a gap between the theoretical objectives of instrument development and the implementation in some test items. Therefore, revisions are necessary to ensure that the cognitive demands in each item truly reflect the characteristics of HOTS. The urgency of this finding is quite high because instruments that appear to align with the curriculum in content may not necessarily be able to accurately measure higher-order thinking skills. If the test items are still at a low cognitive level, the measurement results obtained can provide a biased picture of the students' abilities. Students might obtain high scores simply because they can remember information, not because they can perform complex thinking processes. This certainly contradicts the research objective of producing a valid and accurate HOTS measurement tool.

The most obvious weaknesses are seen in items 2, 4, 8, 11, 12, and 14. The validators assessed that those items are still too close to the cognitive levels C2 or C3. However, the target for instrument development in this research is at the C4, C5, and C6 levels according to the product specifications in the proposal. Therefore, these findings indicate that the revisions are not only technical but also conceptual in nature to ensure that the instrument remains aligned with the theoretical framework of HOTS development. On the other hand, the validators' assessment shows that the language and instructions aspects of the instrument received very good scores. This indicates that the instrument has been developed with consideration of the readability level appropriate for the characteristics of elementary school students. The language used is considered clear, simple, and easy to understand by fourth-grade students. This advantage is an important aspect because HOTS questions at the elementary school level should not be structured with overly complex language. In the context of education in island regions, the use of local context is of significant importance. Stimuli that are close to the students' daily lives will help them connect knowledge with real experiences. This condition allows students to focus more on the thinking process rather than trying to understand a situation that is foreign to them. Therefore, improvements on the weak items need to be directed not only at enhancing cognitive levels but also at strengthening the context relevant to students' lives. If compared to previous research referenced in the proposal, the results of this study show a number of similarities. This study is in line with the research by Sinaga and Yusuf (2023), which states that the HOTS instrument can be declared valid based on the Aiken’s V index, but still requires further refinement before being widely used. In addition, this research is also in line with various development studies based on the 4D model that place expert validation as an important stage before field trials are conducted. However, this

research has several notable differences compared to previous studies. In this study, content validity is not only understood as the general appropriateness of the material but is also specifically examined from the aspects of HOTS level accuracy, relevance to the island context, and readability for fourth-grade elementary school students. This approach provides a more comprehensive assessment of the instrument's quality. Thus, the content validity in this study not only assesses the appropriateness of the content but also the suitability of the instrument with the characteristics of the learners and the learning context. If compared to the research by Rini and Rufi'i (2023) also mentioned in the proposal, the difference in research focus is quite clear. The research places more emphasis on the analysis of IPAS questions that are already available in schools. The results of the study indicate that the validity of the questions is still low, the reliability is in the moderate category, and the difficulty level of the questions tends to be easy. On the contrary, this research not only analyzes existing instruments but also systematically develops new instruments thru the 4D model stages from define to develop. The similarity between the two studies lies in the finding that IPAS assessments at the elementary school level still face various quality issues. Both studies show that the available evaluation instruments are not yet fully capable of measuring students' higher-order thinking skills. However, this research goes further by offering a solution thru the systematic development of HOTS instruments. Thus, this research not only identifies the problem but also strives to provide more concrete alternative solutions. The findings regarding content validity also reinforce the initial issues identified in the research proposal. The issue is related to the dominance of LOTS assessments and the unavailability of truly valid, reliable, and practical HOTS UAS instruments, especially for the context of island regions. Therefore, it can be understood as evidence that the instrument development process has been progressing in the expected direction. Nevertheless, the quality of the instrument's content still needs to be improved thru a revision process. Based on the results of the analysis and discussion, the content validity test of the UAS HOTS IPAS instrument for 4th-grade elementary school shows that the instrument has an average Aiken's V value of 0.756, which generally falls into the fairly valid category. Out of the total fifteen developed items, eight items fall into the very valid category, one item falls into the fairly valid category, and six items fall into the less valid category. The main strengths of the instrument lie in the relevance of the material, clarity of language, and completeness of the instructions. Meanwhile, the main weaknesses of the instrument lie in the consistency of HOTS level demands and the contextual stimulus strength in several items. Based on these conditions, the instrument is deemed suitable to proceed to the next development stage by revising the items that are still weak. Revisions should primarily focus on enhancing cognitive demands so that each item truly measures abilities at levels C4, C5, and C6. In addition, strengthening contextual stimuli is also necessary to make the questions more relevant to students' learning experiences. With these improvements, the instrument is expected to function as a more valid and accurate HOTS measurement tool in accordance with the goals of the Merdeka Curriculum and the learning needs in elementary schools, especially in island regions.

Results of the Construct Validity Test of the UAS HOTS IPAS Instrument for 4th Grade Elementary School Using Confirmatory Factor Analysis (CFA)

The construct validity test in this study was conducted after the instrument passed the content validation stage and the initial revision process based on expert input. In accordance with the design established in the research proposal, the construct validity test was conducted at the extensive trial stage involving a relatively large sample size. The analysis used is Confirmatory Factor Analysis (CFA) to test whether the theoretical structure designed in the instrument is truly supported by empirical data obtained from the field. Thru CFA, this study aims to confirm the suitability of the Higher Order Thinking Skills (HOTS) measurement model constructed based on

three cognitive dimensions, namely C4 (analyzing), C5 (evaluating), and C6 (determining solutions or creating). Based on the extensive trial data used in the workbook, the number of respondents analyzed was 120 students from four elementary schools, with a distribution of 30 students in each school. All students completed the instrument in full, allowing all responses to be utilized in the CFA analysis process without any data being eliminated. Data collection at this stage is specifically aimed at obtaining student responses necessary for testing the construct validity of the instrument. The response data were then coded in the form of 1 for correct answers and 0 for incorrect answers, according to the response matrix format prepared as the basis for data processing in CFA analysis. The research results show that the construct validity of the UAS HOTS IPAS test instrument for fourth-grade elementary school students generally meets the expected criteria based on Confirmatory Factor Analysis (CFA). These findings directly address whether the theoretical model designed by the researchers receives empirical support from the data obtained from extensive trials. The theoretical structure consisting of three factors, namely C4, C5, and C6, has been proven to be adequately represented by empirical data. Thus, the conceptual model used in the development of the instrument is not only theoretical but also gains statistical legitimacy thru construct analysis. The main findings from the CFA analysis indicate that all the key goodness of fit indices have met the required criteria. A p-value of 0.067, χ^2/df of 1.297, RMSEA of 0.031, CFI of 0.958, TLI of 0.949, and GFI of 0.936 indicate that the proposed measurement model has a good fit with the empirical data. These indices collectively indicate that the designed factor structure can adequately explain the relationships between items in the instrument. Thus, the measurement model used can be deemed suitable to represent the measured HOTS construct. The feasibility of this model has important methodological implications for the development of assessment instruments in schools. In educational practice, many instruments are used directly after expert validation without empirical testing of their construct structure. This condition has the potential to produce instruments that are theoretically sound but do not fully reflect the construct being measured. Therefore, the CFA findings in this study indicate that the instrument has progressed beyond the content feasibility stage to a stronger construct validity proof stage.

The next finding shows that all items have a loading factor value above 0.30 and are statistically significant. This condition indicates that each item contributes to the HOTS construct measured in the instrument. Methodologically, these results align with the decision criteria that state items with a loading value ≥ 0.30 and significant can be declared constructively valid. Thus, there are no items that completely fail to support the constructed measurement model. In the context of instrument development, the condition does not require item elimination, but rather should be addressed thru minor revisions. Revisions are necessary to improve the accuracy of the indicators against the measured construct. If items with relatively low loading are retained without improvement, the convergent quality of the factor may become suboptimal. Therefore, the decision to make minor revisions indicates that the instrument development process is conducted carefully and based on empirical evidence. The next finding relates to convergent validity at the construct level. The analysis results show that the Composite Reliability (CR) values of all factors have met the established criteria. This indicates that the indicators within each construct have adequate internal consistency in measuring the intended HOTS dimensions. Thus, from the perspective of construct reliability, the instrument can be stated to have a good level of consistency. However, the Average Variance Extracted (AVE) values show variation in each factor. Factor C6 has fully met the AVE criteria, while factors C4 and C5 are still in the marginal category. This condition indicates that the indicators in factor C6 have a stronger ability to explain the variance of its construct. Conversely, the indicators in factors C4 and C5 still require refinement to optimize their contribution to the construct. If compared to previous research referenced in the proposal, the

results of this study show several significant similarities. These findings are in line with the research by Sudiryo, Hartinah, and Susongko (2024), which emphasizes the importance of proving construct validity in the development of HOTS IPAS assessments in elementary schools. Both studies have the same orientation, which is to ensure that the instruments are not only curriculum-relevant but also have a strong psychometric foundation. In addition, the use of CFA in this study is also consistent with the approach theoretically described in the proposal. However, this research also has characteristics that distinguish it from several previous studies. The instrument developed is specifically designed for the UAS HOTS IPAS test for 4th-grade elementary school students in the context of island regions, incorporating local elements into the question design. In addition, this study uses a three-factor CFA model that explicitly maps items at the C4, C5, and C6 levels. Unlike several previous studies that used the Rasch approach or stopped at content validity and reliability, this research continues the analysis to CFA, CR, AVE, and discriminant validity, thus providing a more comprehensive psychometric proof. Based on the analysis and discussion results, the construct validity test of the UAS HOTS IPAS grade IV elementary school test instrument thru CFA shows that the three-factor model consisting of C4, C5, and C6 falls into the fit or fairly fit category with minor revision needs. All the main model fit indices meet the required criteria, namely a p-value of 0.067, χ^2/df of 1.297, RMSEA of 0.031, CFI of 0.958, TLI of 0.949, and GFI of 0.936. Of the 15 items analyzed, 13 items were declared valid in terms of construct, while two items required minor revisions with no items needing to be eliminated. The CR values of all factors have met the criteria, while the AVE values indicate that factors C4 and C5 are still in the marginal category and factor C6 has fully met the requirements. Overall, the instrument can be stated to have good construct validity and is suitable to proceed to the reliability analysis stage and item quality evaluation, with limited improvements on items with relatively low loading factors.

The Reliability Level of the UAS HOTS IPAS Test Instrument for 4th Grade Elementary School Based on Classical Test Theory Analysis

The reliability test of the UAS HOTS IPAS instrument for 4th-grade elementary school was conducted in the limited trial or initial empirical trial stage. This stage aims to obtain student response data necessary for calculating the internal consistency level of the instrument. The data is used to determine the extent to which the items in the instrument provide stable and consistent results when used on the same group of respondents. The process of testing reliability is an important part of the instrument development stage so that the resulting instrument has adequate measurement quality. Thru this stage, researchers can ensure that the developed instrument is suitable for measuring learning outcomes. The test administration guidelines in this study stipulate that each student's answer is coded dichotomously. Correct answers are scored 1, while incorrect answers are scored 0. This coding system is used because the reliability of the instrument is calculated using the Kuder-Richardson 20 (KR-20) formula. This approach allows researchers to measure the internal consistency of the instrument based on the proportion of correct and incorrect answers for each item. Thus, data processing is carried out systematically to produce an accurate reliability value. Based on the data available in the workbook, the number of respondents involved in the reliability test is 120 students. The instrument analyzed consists of 15 items that have undergone a previous selection stage. The results of the calculation recap show that the value of Σpq is 3.600, while the variance of the total score (St^2) is 12.550. Based on these calculations, the obtained KR-20 reliability coefficient is 0.764. This value indicates that the instrument falls into the fairly good category, thus can generally be considered quite reliable, although it still requires minor revisions in some parts of the instrument. The research results show that the reliability of the UAS HOTS IPAS test instrument for 4th-grade elementary school students is in the fairly good category, with a KR-20 coefficient of 0.764. The value indicates that the developed instrument has

an adequate level of internal consistency based on classical test theory analysis. These findings are related to the reliability level of the instrument produced thru the development process. Thus, the instrument developed not only meets the aspect of validity but also demonstrates a fairly good measurement stability. In the context of instrument development research, reliability is an important indicator that shows the extent to which a test produces consistent scores. Reliability is not only related to the stability of measurement results but also reflects the alignment between items in measuring the same ability construct. This research proposal emphasizes that reliability is the level of consistency of measurement results obtained from an instrument. Therefore, for instruments in the form of multiple-choice questions as in this study, the approach used is the KR-20 reliability coefficient. A reliability coefficient of 0.764 has significant methodological implications in instrument development research. The value indicates that the instrument produced has surpassed the initial draft stage, which was still weak, and has reached an adequate level of stability. Thus, the instrument can be considered sufficiently mature for use in the process of measuring learning outcomes. This confirms that the development process has produced a relatively reliable evaluation tool. Practically, a reliability value above 0.70 indicates that the test results have a sufficient level of stability to differentiate students' learning achievements. In the context of assessment in the 4th grade for the IPAS subject, this condition is very important because the test is used to measure learning outcomes at the end of the semester. Reliable instruments allow teachers to obtain a more accurate picture of students' abilities. Thus, the assessment results obtained can serve as a strong basis for making learning decisions. Another relevant finding in this analysis is that most of the test items have a difficulty level in the moderate category. This condition positively contributes to the improvement of the instrument's reliability. Items that are neither too easy nor too difficult tend to produce greater score variation among test participants. The variation in scores ultimately strengthens the internal consistency of the instrument.

Some items with a proportion of correct responses close to 0.50 yield relatively high pq values. A high pq value indicates that the item has a good ability to contribute to the variance of the total score. This contribution is important because the reliability of the test is greatly influenced by the magnitude of the score variance produced. Therefore, the presence of items with moderate difficulty levels becomes a factor that supports achieving a fairly good reliability. Conversely, items that are too easy or too difficult tend to contribute less to the internal consistency of the test. Items that are too easy cause most students to answer correctly, resulting in a small score variation. Meanwhile, items that are too difficult cause most students to answer incorrectly, which also reduces score variation. Therefore, items with such extreme characteristics are recommended to be revised during the instrument refinement stage. The distribution of total scores obtained in this study also provides important information regarding the performance of the instrument. Scores ranging from 0 to 15 indicate that the test is capable of producing a fairly wide score range. A wide score range suggests that the instrument has good sensitivity in distinguishing students' abilities. This is one of the indicators that the instrument is able to capture variations in test-takers' abilities. A relatively even score distribution also strengthens the interpretation of the obtained reliability coefficient. Instruments that produce scores highly concentrated at one point usually have lower reliability. In this study, the fairly varied score distribution indicates that the instrument is able to differentiate students' ability levels more effectively. This condition reinforces the meaning of the KR-20 reliability coefficient of 0.764. If compared to previous research referenced in the proposal, the results of this study show a number of similarities. The research by Rini and Rufi'i (2023) also shows that IPAS questions at the elementary school level can be analyzed using the classical test theory approach. In the study, the reliability of the instrument was also used as one of the important

indicators in assessing the quality of the questions. This similarity indicates that the approach used in this study is in line with previous research practices. However, there are also quite significant differences between this study and previous research. In the study by Rini and Rufi'i (2023), the reliability of the instrument was reported to still be in the moderate category. Meanwhile, the results of this study show a reliability coefficient that has exceeded the practical limit of 0.70. Thus, the instrument developed in this study can be said to be better prepared for use in the context of summative assessment. This research also shares similarities with the explanation in the proposal that emphasizes the importance of reliability testing at the development stage in development research. This stage aims to assess the internal consistency between items before the instrument is finalized as the end product. The equation shows that the instrument development process follows a systematic methodological logic. In this approach, validity and reliability are viewed as two complementary aspects in determining the quality of the instrument.

If compared to the research by Sinaga and Yusuf (2023) mentioned in the proposal, this study has similarities in the orientation of developing HOTS-based instruments. Both studies place high-order thinking skills as the main focus in question formulation. However, this research has more specific characteristics because it is developed for the context of UAS IPAS for 4th-grade elementary school in island regions. The context provides a different practical dimension in the development of the instrument. Thus, the discussion shows that the UAS HOTS IPAS test instrument for 4th-grade elementary school has achieved a fairly good level of reliability. A reliability coefficient of 0.764 indicates that the instrument has met the minimum reliability threshold required in educational measurement. This condition indicates that the instrument has been sufficiently consistent for use in measuring student learning outcomes. However, the analysis results also indicate the need for improvements to several items. Based on the analysis and discussion that have been conducted, the reliability of the UAS HOTS IPAS class IV SD test instrument, analyzed using classical test theory thru the KR-20 formula, shows a coefficient of 0.764. The value places the instrument in the fairly good category, thus it can be considered sufficiently reliable for use in measuring the learning outcomes of IPAS in the fourth grade. This level of reliability is supported by a fairly varied distribution of total scores and the dominance of items with moderate difficulty levels. However, since the reliability coefficient has not yet reached the very high category, the instrument still requires minor revisions, especially on items that are too easy or too difficult, so that the internal consistency of the test can be further improved.

Quality of UAS HOTS IPAS Class IV Elementary School Test Instrument Items Based on Analysis of Discrimination Power, Difficulty Level, and Distractor Effectiveness

The analysis of item quality in this study was conducted on 15 multiple-choice items in the UAS HOTS IPAS test instrument for 4th-grade elementary school students. The analysis process is based on three main indicators, namely the difficulty level (P), discrimination index (D), and effectiveness of distractors as formulated in the research proposal. The level of difficulty is used to identify the proportion of students who can answer the question correctly, thereby illustrating the ease or difficulty of a question. The discrimination index functions to assess a question's ability to differentiate between students with high and low abilities. Meanwhile, the effectiveness of distractors is used to assess whether the incorrect answer options can optimally divert the choices of students who have not mastered the material. The number of participants involved in the empirical test of this item analysis was 120 students. The final recapitalization results show that out of the total 15 items analyzed, 10 items are deemed worthy of retention, while the other 5 items require revision. No items were found that needed to be eliminated from the instrument. The findings indicate that, in general, the developed instrument has met the criteria for good item

quality. However, some items still require limited improvements to make the resulting measurement quality more optimal. The research results show that the quality of the UAS HOTS IPAS test instrument items for 4th-grade elementary school students is generally in the good category. This assessment is based on the fact that the majority of items have met the empirical indicators set in the test quality analysis. Out of the total 15 items analyzed, 10 items were deemed worthy of retention, while 5 items required limited revision. There are no items that need to be eliminated from the instrument. This finding indicates that the produced product has surpassed the initial design stage and has developed into an instrument with a sufficiently strong empirical foundation. The second important finding relates to the quality of the item discrimination index. The analysis results show that, in general, the discrimination index falls into the good category. A total of 10 items are classified as having good discrimination, and 4 items have very good discrimination. Only one item falls into the sufficient category, which is item 9. This composition shows that most items have functioned well in distinguishing students' abilities.

Discrimination index is one of the main indicators of item quality in learning evaluation. The research proposal emphasizes that a high discrimination index indicates the item's ability to identify differences in ability between high-achieving and low-achieving students. Items with good discrimination power will provide more accurate information regarding students' mastery of the material. This information is very important for teachers in interpreting learning outcomes. Thus, the quality of good discrimination power becomes an important indicator of measurement validity. However, item 9, which has a moderate discrimination index, still requires special attention. These findings indicate that the item's ability to differentiate students is still not optimal. Several factors may contribute to this condition, such as unclear stimuli or questions that are not challenging enough. In addition, the quality of the answer options can also affect the effectiveness of item discrimination. Therefore, revisions are necessary for the item to function more optimally. Interestingly, the issue with item 9 is not related to the level of difficulty or the effectiveness of the distractors. The indicator that is the weak point is only in the aspect of discrimination power. This shows that the quality of the item is not always determined by the number of problematic indicators. One suboptimal indicator alone can reduce the overall quality of the item. This finding emphasizes the importance of comprehensive item analysis in the development of evaluation instruments. The third finding in this study relates to the effectiveness of distractors in multiple-choice questions. The analysis results show that 11 items have all distractors functioning well. Meanwhile, there are 4 items that still have one ineffective distractor. Those items are items 1, 10, 13, and 15. This condition indicates that, in general, the quality of the answer options is quite good, although there are still some parts that need improvement. The urgency of the quality of distractors is often overlooked in the practice of question formulation in schools. Many question writers focus more on creating the correct answers, while the wrong options are considered mere complements. However, in classical test theory, the quality of distractors directly affects the level of difficulty and item discrimination. Distractors that are illogical or too weak can make the questions easy to guess. This condition ultimately reduces the quality of the measurement produced by the test. Some empirical findings in this study provide a concrete picture of the issue. In item 10, for example, option C was only chosen by one student. The same happened in item 15, when option A was only chosen by one student. In item 13, option D was chosen by 4.17% of the students, slightly below the established effectiveness threshold. This data shows that some distractors have not been able to attract the attention of students who do not understand the material well. The findings regarding ineffective distractors provide very clear directions for revision. Improvements do not need to be made by replacing the entire question item. The revision should focus on reorganizing the answer options to make them more logical and contextual. Distractors need to be designed based on

common misconceptions that students may have. In this way, each answer option can function optimally in measuring students' understanding.

If all the findings of this research are analyzed comprehensively, it can be concluded that the quality of the instruments is generally in the good category. However, some items still require refinement to make their measurement functions sharper. The five revised items do not indicate a failure in the development process. On the contrary, the revision actually shows that the instrument development was carried out systematically and based on empirical evidence. In development research, such improvement processes are a normal part and actually serve as indicators of the quality of the research process. The results of this study are also related to several previous studies. If compared to the research by Rini and Rufi'i (2023), there are similarities in the findings regarding the importance of item analysis in the evaluation of IPAS learning in elementary schools. Both studies show that the item discrimination index is generally in the good category. However, Rini and Rufi'i's research found that most of the questions were in the easy category. Meanwhile, this study shows a more proportional distribution of difficulty as it is dominated by the moderate category. A comparison can also be made with the research by Syahrul et al. (2025) conducted in the context of Selayar Islands Regency. The research emphasizes the importance of item analysis training based on classical test theory for teachers. The results show that the training is capable of enhancing teachers' understanding of validity, difficulty level, discrimination power, and distractor effectiveness. The similarity with this research lies in the emphasis on the importance of psychometric literacy for teachers. The difference is that this research does not stop at the training aspect but also produces HOTS instruments that are empirically analyzed. If compared to the research by Sudiryo, Hartinah, and Susongko (2024), there is a similarity in the focus on developing HOTS IPAS assessments for elementary schools. However, the methodological approaches used are different. Sudiryo et al. used the Rasch model in instrument analysis. Meanwhile, this study uses a classical test theory approach combined with confirmatory analysis and practicality testing. This approach yields more direct information regarding which items need to be retained or revised. Comparison with the research by Sinaga and Yusuf (2023) also shows differences in the scope of instrument testing. The study emphasizes the content validity of the HOTS instrument. However, its analysis does not yet cover construct validity and practicality testing. This research addresses those shortcomings by conducting a more comprehensive evaluation. The instrument was not only tested in terms of content but also construct, reliability, item quality, and practicality. Overall, the findings have significant implications for the quality of the final product of the instrument. The research proposal emphasizes that the final instrument will only include items that meet various quality criteria. These criteria include content validity, construct validity, proportional difficulty level, good discrimination power, and high reliability. Thus, item analysis serves as the basis for decision-making regarding the suitability of each item. This process ensures that the resulting instrument truly has optimal measurement quality. Based on the analysis and discussion results, the quality of the UAS HOTS IPAS test instrument items for 4th-grade elementary school students is generally in the good category. Of the 15 items analyzed, 14 items have a moderate level of difficulty and 1 item is classified as easy. In terms of discrimination power, 14 items fall into the good to very good category, while 1 item falls into the sufficient category. From the aspect of distractor effectiveness, 11 items had all their distractors functioning, while 4 items still had one ineffective distractor. Overall, 10 items were retained and 5 items were revised without any items being discarded. These findings indicate that the instrument has adequate empirical quality, although limited revisions are still necessary to enhance measurement precision.

The Practicality Level of the UAS HOTS IPAS Test Instrument for 4th Grade Elementary School Based on the Implementation Test Results and Teacher and Student Responses

The practicality test of the UAS HOTS IPAS instrument for 4th-grade elementary school is conducted to assess the ease of use, ease of understanding, and feasibility of implementing the instrument in the UAS execution in elementary schools in the island region. This testing is intended to ensure that the developed instrument is not only conceptually valid but also easy to implement by users in the field. According to the design in the research proposal, the sources of practicality data are obtained from three main components, namely the observation of test implementation, teacher response questionnaires, and student response questionnaires. The general practicality indicators used in the proposal include the ease of understanding instructions, clarity of the language used, coherence between stimuli and questions, ease of scoring process, and the appropriateness of the time allocation provided for taking the test. The practicality measurement instrument used is also structured in accordance with the format of the instrument that has been previously designed. The practicality recap sheet contains three main components that form the basis of the assessment, namely the recap of implementation observation results, the recap of teacher response questionnaires, and the recap of student response questionnaires. These three components are then integrated to obtain the overall average practicality score of the instrument. In addition, the decision-making format used explicitly requires the reporting of several indicators, namely the percentage of test implementation, the percentage of teacher responses, the percentage of student responses, and the average total practicality score of the instrument. The number of subjects involved in the practicality test consists of three observers, three teacher respondents, and thirty student respondents. The data recapitulation results show that the average percentage of test implementation reached 91.7%, while the average teacher response was 90.0%. Meanwhile, the average student response was recorded at 88.4%. Thus, the total average practicality score of the instrument reached 90.0%, categorizing the developed test instrument as very practical for use in the implementation of UAS in elementary schools in the island region. The results of this study indicate that the practicality level of the UAS HOTS IPAS test instrument for fourth-grade elementary school students falls into the very practical category. The average total practicality obtained reached 90.0%, indicating that the instrument is not only valid and reliable but also easy to implement. The developed product can be positively accepted by users in the field. The aspect of practicality becomes an important part of the expected final product characteristics, in line with validity and reliability. All components of the instrument's practicality fall into the very practical category, including feasibility, teacher response, and student response. The feasibility percentage of 91.7%, teacher response of 90.0%, and student response of 88.4% show very strong consistency. This confirms that the instrument is not only considered practical by one group but is uniformly accepted by all user groups. That consistency is important in development research because the product may be deemed easy by researchers, but it may not necessarily be practical for teachers or students. The practicality of this instrument becomes significant when linked to the initial issues raised in the research proposal. The proposal emphasizes that one of the main issues in the field is the lack of standardized, valid, reliable, and practical HOTS test instruments, especially in island regions. Thus, the finding regarding the very high practicality not only complements the findings but also demonstrates that this research successfully addresses the real needs in elementary schools. A valid but difficult-to-use instrument will not provide significant benefits for either teachers or students. Teachers' responses to the instrument also fall into the very practical category, with an average of 90.0%. Teachers assessed that the usage instructions were clear, the test format was easy to use, and the language of the questions was appropriate for the abilities of fourth-grade students. The answer keys and scoring guidelines were considered easy to apply, while the instrument effectively helped assess students' higher-order thinking skills. These findings indicate that the

product not only meets theoretical aspects but also operational ones. The urgency of the teacher response findings is very high, because teachers are the primary users of the instrument. Teachers play the role of translating curriculum policies into evaluation practices, so their perception of ease of use becomes crucial. Instruments that are well-received by teachers are highly likely to be used sustainably.

This is in line with the research proposal that emphasizes practicality as one of the main criteria for product feasibility. Student responses also showed very positive results, with an average of 88.4%. Out of 30 students, 26 students rated the instrument as very practical, while 4 students rated it as practical. This shows that the instructions are easy to understand, the language of the questions is clear, and the stimuli in the form of pictures or stories help students in understanding the questions. This finding is important because it shows that HOTS instruments remain accessible to elementary school students without losing their high-level thinking content. The practicality from the students' perspective shows a successful balance between the complexity of thinking and the students' cognitive abilities. The instrument demands high-level thinking processes thru meaningful contexts, while still considering accessibility for elementary school children. These findings reinforce the relevance of the instrument with the HOTS question design principles, which are not solely based on linguistic difficulty or reading length. Thus, the questions remain cognitively challenging but do not confuse the students. Although the quantitative scores indicate very high practicality, the qualitative data reveal the need for minor revisions. Observers highlighted the contrast in images, teachers suggested simplifying the instructions, and students indicated that some story stimuli still felt lengthy, affecting time adequacy. These minor revisions are important to enhance the user experience without compromising the quality of the test. This demonstrates the development's responsiveness to real field feedback. The proposed minor revisions focus on technical aspects that can affect readability and comfort. Improving image contrast, example instructions, and time allocation for long stimuli will enhance the clarity and effectiveness of instrument use. Although it may seem minor, its impact on the quality of the experience for students and teachers is significant. This aspect emphasizes that the development of educational products should be carried out iteratively and based on field evidence. In the context of the proposal, these practical findings support the expected characteristics of the final product, which are practical, relevant, and contextual. The local island context integrated into the instrument actually helps students' understanding, rather than causing confusion. This shows that the instrument successfully combines HOTS demands with the reality of learning in elementary schools in island regions. The findings of positive student responses support that assumption. If compared to previous research, this finding is similar to Fitria, Wijaya, and Danial (2020) which shows that HOTS-based products can reach the very practical category. The similarity lies in the ease of use of the product by the users. However, there is a significant difference because Fitria et al.'s research focuses on LKPD, while this research focuses on the UAS test instrument. Therefore, the practicality in this study is related to test administration and scoring. The results of this study are also in line with the findings of Raras, Siswanto, and Wijayanti (2024) regarding the contribution of HOTS devices in the IPAS learning of the fourth grade. The similarity lies in the development orientation that places HOTS as the main focus. The difference lies in the focus of the study; Raras et al. emphasized the effectiveness of LKPD on learning outcomes, whereas this research emphasizes the practicality of the test instrument as a summative evaluation tool. This emphasizes the relevance of the context in which the product is used.

4. CONCLUSION

This research shows that the development of the Semester Final Examination (UAS) test instrument based on Higher Order Thinking Skills (HOTS) for the 4th-grade IPAS subject in elementary schools in the archipelago region successfully produced instruments that are valid, reliable, practical, and contextual. The development was carried out thru the 4D model systematically, starting from needs analysis to product testing, so that the resulting instrument is in accordance with the characteristics of elementary school students and the learning context in the island regions. The test results show that the instrument has good content validity and construct validity, adequate reliability, and item quality that can more accurately measure students' higher-order thinking skills. The theoretical contribution of this research lies in the development of a HOTS IPAS assessment model for elementary schools that integrates content validity, construct validity, reliability, item quality analysis, and practicality within a comprehensive instrument development framework. This research also reinforces the study that HOTS assessments for elementary school students can be developed contextually by considering the cognitive development stages of students and the characteristics of the learning area. In addition, this research provides a practical contribution in the form of HOTS evaluation instruments that can be used by teachers as an alternative assessment to support the implementation of the Merdeka Curriculum, particularly in elementary schools in island regions.

5. REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

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

Conklin, W. (2012). *Higher-order thinking skills to develop 21st century learners*. Shell Education.

Fitria, D., Wijaya, A., & Danial, M. (2020). Pengembangan LKPD berbasis HOTS menggunakan model 4D untuk meningkatkan kemampuan berpikir tingkat tinggi siswa. *Jurnal Pendidikan dan Pembelajaran*.

Hattarina, S., et al. (2022). Implementasi Kurikulum Merdeka dalam pembelajaran IPAS di sekolah dasar. *Jurnal Pendidikan Dasar*.

Hulaipah, N., Syukri, M., & Indraswati, D. (2023). Analisis kesulitan guru dalam menyusun soal HOTS di sekolah dasar. *Jurnal Ilmiah Pendidikan*.

Kemendikbud. (2021). *Panduan pembelajaran dan asesmen kurikulum merdeka*. Kementerian Pendidikan dan Kebudayaan.

Lamiah. (2025). Analisis kemampuan siswa dalam menyelesaikan soal HOTS pada mata pelajaran IPAS kelas IV SD. *Jurnal Pendidikan Dasar*.

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

Nurgiyantoro, B. (2018). *Penilaian pembelajaran bahasa berbasis kompetensi*. BPFE Yogyakarta.

Piaget, J. (1952). *The origins of intelligence in children*. International Universities Press.

Raras, P., Siswanto, J., & Wijayanti, A. (2024). Pengembangan LKPD berbasis HOTS untuk meningkatkan hasil belajar IPAS siswa SD. *Jurnal Pendidikan dan Pembelajaran*.

Rini, D., & Rufi'i. (2023). Analisis butir soal IPAS kelas IV SD. *Jurnal Evaluasi Pendidikan*.

Saputra, A., Asrin, & Novitasari. (2024). Implementasi pembelajaran berbasis HOTS di sekolah dasar. *Jurnal Pendidikan Dasar Indonesia*.

Sinaga, R., & Yusuf, M. (2023). Pengembangan instrumen tes berbasis HOTS menggunakan model ADDIE pada materi asam basa. *Jurnal Penelitian Pendidikan Sains*.

Sudaryono. (2012). *Dasar-dasar evaluasi pembelajaran*. Graha Ilmu.

Sudiryo, A., Hartinah, S., & Susongko, P. (2024). Pengembangan asesmen HOTS berbasis model Rasch pada mata pelajaran IPAS SD. *Jurnal Evaluasi Pendidikan Indonesia*.

Sugiyono. (2019). *Metode penelitian pendidikan (kuantitatif, kualitatif, dan R&D)*. Alfabeta.

Sukardi. (2015). *Evaluasi pendidikan: Prinsip dan operasionalnya*. Bumi Aksara.

Tanujaya, B., Mumu, J., & Margono, G. (2017). Higher order thinking skills berdasarkan revisi Taksonomi Bloom. *Jurnal Pendidikan dan Pembelajaran*.

Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Indiana University.

Widodo, T., & Kadarwati, S. (2013). Higher order thinking berbasis pemecahan masalah untuk meningkatkan hasil belajar. *Jurnal Pendidikan*.

Zainal, A. (2012). *Evaluasi pembelajaran*. Remaja Rosdakarya.

Zubaidah, S. (2018). Mengenal 4C: Learning and innovation skills untuk menghadapi abad 21. *Seminar Nasional Pendidikan Biologi*.