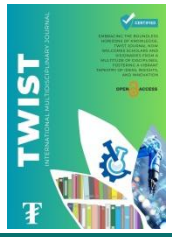
DOI 10.5281/twist.10049652#549

TWIST, 2026, Vol. 21, Issue 1, pp. 45-56

TWIST

Journal homepage: www.twistjournal.net



Validity and Reliability of HOTS Instruments for Early Childhood

A Preliminary Study

Danang Prastyo*
State University of Surabaya, Indonesia
ORCID: https://orcid.org/0000-0002-6822-3257
[**Corresponding author*]

Gunarti Dwi Lestari
State University of Surabaya, Indonesia
ORCID: https://orcid.org/0000-0003-4688-844X

Andi Kristanto
State University of Surabaya, Indonesia
ORCID: https://orcid.org/0000-0002-8127-2707

Abstract
This study aims to develop and examine the validity and reliability of an assessment instrument for measuring Higher Order Thinking Skills (HOTS) in early childhood. The lack of instruments tailored to the developmental characteristics of young children remains a gap in the field of education, as most existing HOTS measurement tools are not specifically designed for early childhood learners. This research employed a quantitative approach within the framework of Research and Development (R&D), focusing on content validity through expert judgment, construct validity, and reliability testing. The study was conducted at TK Al-Quran Al-Hakim, located in Burneh Village, Bangkalan Regency, with a sample of 30 children in group B, approximately 6 years old. The results indicated that the content validity assessed by three experts yielded a score of 86.1%, categorized as "good." Construct validity testing showed that 10 out of 15 items were valid (r calculated > 0.361). Furthermore, the reliability test produced a Cronbach's Alpha value of 0.838, indicating high internal consistency. Therefore, the developed HOTS instrument is considered valid and reliable, and it can be effectively used to assess higher-order thinking skills in early childhood education.

Keywords
Higher Order Thinking Skills (HOTS), early childhood, content validity, construct validity, reliability, assessment instrument

INTRODUCTION

Higher Order Thinking Skills (HOTS) have become an integral component of modern education systems that emphasize meaningful, contextual, and learner-centered instruction (Zebua, 2024). HOTS refer to cognitive abilities that go beyond mere recall of information, encompassing skills such as analyzing, evaluating, and creating based on existing knowledge (Nisa et al., 2018). In the revised version of Bloom's taxonomy by Anderson and Krathwohl (2001) as cited in (Wewe & Wangge, 2021), HOTS are situated at the three highest cognitive levels: analyzing, evaluating, and creating (Kunduracı, H., Yaralı, K., & Kaynak, 2024). These skills are essential to develop from an early age, as this period is marked by rapid brain growth and development (Paudpedia, 2022). By designing a learning environment that stimulates reflective thinking, young children can be guided to gradually develop HOTS in accordance with their developmental stages.

Theoretically, the development of HOTS in early childhood is grounded in the understanding that, although children are still in the preoperational stage according to (Piaget, 1952) , a stage characterized by the emergence of symbolic representation and imagination—they nevertheless possess fundamental capacities to exhibit early signs of critical and reflective thinking within concrete contexts. Young children can demonstrate basic cause-and-effect

reasoning, make choices based on simple arguments, and generate new ideas during play activities (Yıldız, C., & Yildiz, 2021). Therefore, HOTS in early childhood should not be equated with those in adolescents or adults, but rather understood as advanced thinking abilities appropriate to the cognitive structure of children during the golden age of development (Sulaiman, 2020).

In the practice of early childhood education in Indonesia, the assessment of children's thinking skills tends to remain descriptive and qualitative, often lacking instruments specifically designed to assess HOTS (Purnasari et al., 2021). However, valid and reliable assessments are crucial for objectively determining the extent of children's cognitive development (Frausel et al., 2020). Unfortunately, to date, few instruments are available that are specifically tailored to measure HOTS in early childhood. Existing tools generally focus on broader developmental domains such as motor, language, and socio-emotional skills, or remain limited to measuring Lower Order Thinking Skills (LOTS), such as remembering or understanding information (Setiawati et al., 2019). This highlights a significant gap between the urgent need for HOTS development and the availability of relevant and trustworthy assessment tools for early childhood education.

The lack of standardized instruments to assess Higher Order Thinking Skills (HOTS) in early childhood represents a serious issue that affects the learning process. Without appropriate assessment tools, educators face challenges in designing instructional strategies that can systematically stimulate HOTS (Li, W., Huang et al., 2023). Moreover, assessment processes that are not grounded in valid instruments risk producing biased data, which may fail to reflect children's actual cognitive abilities. Therefore, the development of valid and reliable HOTS instruments has become an urgent need within early childhood education systems that prioritize the development of children's thinking potential from an early age.

Previous studies related to the development of HOTS instruments for early childhood education remain limited. For instance, (Marada et al., 2021) developed a HOTS-based instrument aimed at fostering critical thinking in biology education; however, the study did not include a quantitative test of construct validity. Another study by (Rodiana & Pahlevi, 2020) developed a HOTS-based assessment instrument for the archival science subject in Office Administration programs, attempting to adapt a thematic-based assessment approach. (Fitriani & Vinayastri, 2022) designed a critical thinking instrument for early childhood learners, but the differentiation between HOTS and LOTS indicators was not conducted systematically, nor was the instrument tested through a confirmatory factor analysis (CFA) approach. These studies demonstrate a growing interest in developing HOTS instruments in early childhood education, but most remain at the exploratory stage. This indicates a need for further research with stronger methodological rigor to examine the psychometric properties of the developed instruments.

In light of the need for HOTS assessment instruments that align with the developmental characteristics of young children, as well as the limited theoretical and empirical exploration in this field, the present study is conducted as an initial investigation focusing on testing the validity and reliability of a HOTS instrument for early childhood. This research is expected to make a significant contribution to the development of assessment tools that can be utilized by early childhood educators, researchers, and policymakers in building a more accurate, holistic, and developmentally appropriate assessment system that supports children's cognitive growth.

To ensure the psychometric quality of the developed instrument, this study employed two types of validity testing: (1) expert judgment validity, which involved evaluations by experts to assess the alignment of each item with indicators of Higher Order Thinking Skills (HOTS) in early childhood; and (2) construct validity, which was analyzed quantitatively using SPSS software to examine the inter-item relationships and the underlying factor structure. Meanwhile, the reliability of the instrument was tested using the internal consistency coefficient, specifically Cronbach's Alpha, to determine the extent to which the items consistently measure the same construct.

The research questions addressed in this study are as follows: (1) What is the expert judgment validity of the developed HOTS instrument for early childhood? (2) What is the construct validity of the developed HOTS instrument for early childhood? and (3) To what extent does the instrument demonstrate internal reliability in measuring higher order thinking skills in early childhood? Focusing on the initial testing of the instrument's quality, this study is limited to the age range of 4 to 6 years and adopts a descriptive quantitative approach with basic statistical analysis. The findings from this preliminary study are expected to serve as a foundation for the development of more comprehensive instruments that can be widely implemented in early childhood education settings.

MATERIALS AND METHODS

This study employed a Research and Development (R&D) approach aimed at developing and testing the validity and reliability of an assessment instrument for Higher Order Thinking Skills (HOTS) in early childhood. The primary focus of this research was to conduct content validity testing, construct validity analysis, and reliability testing as part of a preliminary study in developing a HOTS assessment instrument that aligns with the developmental characteristics of young children. A quantitative approach was utilized, supported by statistical analyses to ensure the validity and reliability of the instrument. These techniques have proven effective in previous studies, such as that by(Christianti, 2024), which reported a construct validity of 0.97 and high reliability in an early childhood literacy assessment instrument.

The research design was based on a simplified version of the development model proposed by Borg dan Gall (2003), focusing only on the following stages: (1) needs assessment and literature review, (2) development of the initial

draft instrument based on HOTS indicators for early childhood, (3) content validation by experts (expert judgment), and (4) initial construct validity and reliability testing through quantitative data analysis.

This study was conducted at TK Al-Quran Al-Hakim, located in Burneh Village, Burneh District, Bangkalan Regency, East Java. The trial subjects in this study were 30 children from Group B, approximately six years old, who served as the main participants in the construct validity and reliability testing of the HOTS instrument. The developed instrument consisted of 15 items designed to measure higher order thinking skills in early childhood based on the indicators of analysis, evaluation, and creation. These indicators were adapted to the cognitive developmental context of early childhood and were grounded in appropriate assessment principles for young learners, such as authentic and observational assessment (Faizah, 2019). The list of test items is presented in the following table.

Table 1 Test Items as HOTS Instrument

No	Test Item	Cognitive Achievement Level
1	Count the number of vehicle parts, then match them with the correct number!	C4
2	Arrange the vehicle miniatures based on the number of passengers they carry, starting from the fewest!	C5
3	Tell a story about your experience riding a vehicle with your family during the school holiday!	C6
4	Mark with an X the vehicle pictures that emit black smoke, then choose the picture showing how to keep the air clean, and stick it on the board!	C5
5	Create 3 traffic signs using recycled cardboard to help cars follow the rules!	C6
6	Find and circle 5 differences between the car pictures, then mention them!	C5
7	Attach triangle, square, and circle shapes to the missing parts of the vehicle picture!	C4
8	Draw your favorite vehicle! Use your imagination freely!	C6
9	Put a check mark V on the pictures showing appropriate behavior in the car, and an X for inappropriate ones. Why is that?	C5
10	Let's classify these miniature cars! Which ones are used to help people? Which are family cars?	C4
11	Create a travel route from recycled materials for a toy car to reach the school!	C6
12	Group the vehicle miniatures based on their stopping places, then name the vehicles!	C5
13	Make a toy car out of cardboard as you like!	C6
14	Decorate your car with colors and images you like to make it look cooler!	C4
15	Present your self-made toy car in front of your friends!	C4

Content validity testing was conducted through expert judgement involving three subject matter experts: Muchamad Arif, S.Pd., M.Pd. (Lecturer at Narotama University), Desika Putri Mardiani, M.Pd., and M. Fahmi Zakariyah, M.Pd. (both are Lecturers at the State University of Surabaya). The experts evaluated each item in the instrument using a Likert scale ranging from 1 to 4, where a score of 1 indicates "not appropriate" and a score of 4 indicates "highly appropriate." The aspects assessed by the experts included: content (alignment of the material with HOTS indicators for early childhood), language (use of communicative and age-appropriate language), and presentation (layout, clarity of instructions, and visual support materials). The results of the expert evaluations were analyzed to obtain the content validity score, which was used to determine whether each item was suitable for further pilot testing. This validity test was conducted using Aiken's V formula, which is commonly used in assessing content validity of psychological and educational instruments (Azwar, 2017).

Table 2 Expert Validation Instrument

VALIDASI INSTRUMEN TES
OLEH AHLI EVALUSI PEMBELAJARAN

Nama	Muchamad Arif, S.Pd., M.Pd
Jabatan	Dosen
Lembaga	Universitas Narotama

A. Petunjuk Pengisian

1. Mohon memberikan penilaian yang sesuai menurut penilaian Bapak/Ibu dengan memberi tanda (√) pada kolom penilaian yang telah disediakan.
2. Jika ada masukan, saran dan perbaikan, mohon menuliskan komentar dan saran perbaikan yang telah disediakan.
3. Ketentuan pilihan dengan melihat pedoman penilaian 1-4 sebagai berikut:
   1 = Sangat kurang
   2 = Kurang
   3 = Baik
   4 = Sangat Baik
4. Atas kesediaan Bapak/Ibu, diucapkan terima kasih.

B. Pertanyaan

Aspek	Indikator	Penilaian			
		1	2	3	4
Isi	Kesesuaian butir soal yang disajikan dengan perkembangan usia peserta didik				V
	Kesesuaian butir soal yang disajikan dengan kebutuhan bahan ajar			V	
	Uraian yang disajikan tidak menimbulkan penafsiran ganda atau salah pengertian			V	
	Butir soal yang disajikan sesuai dengan ranah kognitif C4-C6 yang di ukur			V	
Kebahasaan	Keterbacaan butir soal			V	
	Kejelasan butir soal				V
	Penggunaan bahasa yang efektif dan efisien				V
	Butir soal yang disampaikan sesuai dengan kaidah EYD			V	
Sajian	Butir soal yang disajikan mencerminkan jabaran yang mendukung pencapaian kompetensi dasar				V
	Butir soal yang disajikan mencangkup materi yang terkandung dalam indikator pencapaian				V
	Butir soal yang disajikan mencangkup materi yang terkandung dalam tujuan pembelajaran			V	
	Butir soal tidak mengandung unsur SARAPPPK (Suku, Agama, Ras, Antar golongan, Pornografi, Politik, Propaganda, dan Kekerasan)				V

Komentar :

Saran :

Kesimpulan Instrumen Tes:
- ☐ Layak Digunakan
- ☐ Tidak Layak Digunakan

Kategori:
- ☐ Sangat Baik
- ☐ Baik
- ☐ Tidak Baik
- ☐ Sangat Tidak Baik

Surabaya, 05 Mei 2025
Validator

MUCHAMAD ARIF, S.Pd., M.Pd

After the expert validators completed the questionnaire, the resulting feasibility scores that met the criteria became part of the media feasibility analysis. The expert questionnaire was used to determine the evaluation outcome of the feasibility test, and the following formula was applied to calculate the instructors' responses:

$$P = \frac{S}{N} \times 100\%$$

Fig. 1 Formula for Media and Content Feasibility Percentage

Description:
P : Ideal Percentage
S : Total Score Obtained
N : Maximum Possible Score

The analysis results from the validation sheet were used to assess the feasibility of the expert validation instrument that had been developed. The interpretation of the results can be seen in the following table:

Table 3 Expert Validation Instrument Feasibility Criteria Based on Percentage Analysis

Score	Classification	Rating
>80	Highly Feasible	4
>60-80	Feasible	3
>20-60	Less Feasible	2
≤20	Not Feasible	1

In addition to content validity, construct validity was also assessed, aiming to determine the extent to which the theoretical construct of Higher Order Thinking Skills (HOTS) is reflected in the items of the developed instrument. Construct validity testing was conducted through inter-item correlation analysis using SPSS software, based on data collected from children in the trial group. The construct validity was calculated to assess the degree of correlation between each item and the total score, indicating the contribution of each item to the measured construct. Given that the number of respondents in the validity test was 30 students, the degrees of freedom (df) were calculated as N - 2, resulting in df = 30 - 2 = 28. At a significance level of $\alpha = 0.05$, the following decision rule was applied: if the calculated r-value (r_{xy}) is greater than the critical r-value from the table, the item is considered valid; conversely, if the calculated r-value is less than the table value, the item is considered invalid. Following the validity test, the next stage was the reliability test, which aimed to determine the level of consistency of the instrument in measuring HOTS in early childhood. Reliability testing was conducted using Cronbach's Alpha method with the assistance of SPSS, which evaluates the internal consistency of all items within the instrument. An instrument is considered to have good reliability if the Cronbach's Alpha coefficient is ≥ 0.60. This value indicates that the items demonstrate high consistency and are capable of producing stable data in repeated measurements. This approach was also employed by (Wantah, 2010) in testing assessment instruments for children with special needs in kindergarten, and by (Utsman, 2013) in his dissertation on the development of instruments for assessing early childhood development achievements.

RESULTS AND DISCUSSION

This study aims to develop and test the quality of an assessment instrument for Higher Order Thinking Skills (HOTS) in early childhood, specifically targeting children aged 4 to 6 years. The primary focus of this study is on three key aspects of instrument quality: expert validity, construct validity, and internal reliability. The instrument testing was conducted in the initial phase to gain an overview of the extent to which the instrument is feasible and reliable for measuring higher-order thinking skills in early childhood within a learning context appropriate to their developmental stage. Therefore, instrument validation is a crucial preliminary step before its broader application in early childhood education (PAUD) settings.

This chapter presents a systematic analysis of each aspect of validity and reliability of the instrument, starting from expert judgment validity, construct validity, and internal reliability testing. The discussion in each section elaborates on the findings obtained and relates them to relevant theories and previous studies, thereby providing a comprehensive understanding of the quality of the developed instrument.

Results of Expert Validation of the Test Instrument

The validation of the test instrument was conducted by three experts in educational assessment with relevant academic backgrounds and professional experience: Muchamad Arif, S.Pd., M.Pd. from Narotama University, and Desika Putri Mardiani, M.Pd., along with M. Fahmi Zakariyah, M.Pd. from the State University of Surabaya. The experts evaluated the instrument based on three key aspects: content validity, language validity, and presentation validity. Assessments were carried out using a Likert scale ranging from 1 to 4, with ratings spanning from "Very Poor" to "Very Good." Each aspect comprised several detailed indicators designed to measure the instrument's appropriateness in the context of teaching and assessment.

Regarding content validity, the experts assessed the alignment of test items with the developmental characteristics of the target students, their relevance to the instructional material, the clarity of item descriptions to prevent ambiguity, and the cognitive levels assessed, corresponding to Bloom's revised taxonomy levels C4 to C6 (analysis, evaluation, and creation). The evaluation yielded a high average score of 85.41%, indicating that the items were generally representative of the competencies intended to be measured. This score also suggests that the instrument was developed with pedagogical considerations, ensuring that the items are suitable for the age group and the content requirements. However, some indicators received scores of 3, which suggests potential for further refinement of item wording to enhance clarity and comprehensibility, particularly to avoid semantic ambiguity. (Arikunto, 2008) defines content validity as the extent to which the items of a test represent the material taught. High content validity confirms that the instrument comprehensively covers essential aspects of the curriculum, thereby enabling accurate measurement of student competencies (Anshari et al., 2024; Ramadhan et al., 2024).

In addition, language validity was given careful attention by the experts, as linguistic clarity is a critical component of instrument validity. This evaluation considered readability, sentence clarity, effective and appropriate language use, and conformity with the Indonesian Spelling System (Ejaan Yang Disempurnakan, EYD). The language aspect achieved an average score of 83.33%, classifying it as highly appropriate and ensuring that students would not encounter difficulties in understanding the test items. Nonetheless, the experts provided some indirect feedback suggesting editorial improvements, such as refining word choice and sentence structure, to further enhance linguistic focus and mitigate any potential ambiguity. This is important because ambiguous language can significantly impact test outcomes, thus it is essential that the instrument is both communicative and easily comprehensible. The study by (Safi'i et al., 2022) underscores the importance of language clarity in the validity of evaluation instruments, demonstrating that clear and grammatically accurate language substantially affects student comprehension of assessment items.

The third aspect, presentation validity, evaluates the layout, clarity of presentation, and the absence of negative elements such as ethnicity, religion, race, inter-group issues (SARA), pornography, politics, propaganda, and violence. The experts awarded the highest average score of 89.58%, indicating that the instrument was developed with strong attention to ethical and aesthetic considerations and aligns with the norms and values prevailing in the educational context. The freedom of the instrument from discriminatory content and negative elements is crucial to maintaining objectivity and fairness in the evaluation process. Furthermore, the systematic presentation of the items, which supports the achievement of basic competencies and learning indicators, makes this instrument not only theoretically valid but also practical for use in actual learning situations. According to (Kaaffah et al., 2021), a well-presented instrument not only enhances aesthetics but also assists learners in better understanding and responding to the test items. Systematic presentation free from negative elements is essential for upholding objectivity and fairness in the evaluation process.

Table 4 Expert Validation Results of the Test Instrument

No	Assessment Aspect	Expert Evaluation			Total per Aspect
		Desika Putri Mardiani	M. Fahmi Zakariyah	Muchamad Arif	
1	Format Isi	81,25%	93,75%	81,25%	85,41%
2	Kebahasaan	81,25%	81,25%	87,5%	83,33%
3	Sajian	87,5%	87,5%	93,75%	89,58%
Overall Aspect Total					**86,1%**

Overall, the combined validity score from the three aspects reached 86.1%, which is above the threshold of 80% according to Arikunto's (2005) criteria. Therefore, this test instrument can be categorized as highly feasible or valid for use in research as well as in the implementation of learning evaluation. This high validity also strengthens the credibility of the instrument in accurately, objectively, and comprehensively measuring students' abilities. The success of this validation indicates that the instrument development process went through a well-considered design phase involving comprehensive scientific and pedagogical considerations.

Furthermore, the absence of suggestions or improvement comments from the experts signifies that the instrument has met the expected quality standards and is ready to be applied without the need for significant revisions. However, as a prudent follow-up in instrument development, it is recommended to conduct an empirical pilot test to determine the instrument's reliability and the real responses from students. This empirical test will be the next crucial step to ensure that the instrument is not only valid in theory but also reliable and effective in real-world application contexts.

Thus, this expert validation not only serves as evidence of the instrument's conceptual quality but also provides a strong foundation to ensure that the test instrument can deliver valid and trustworthy measurement results in the context of learning evaluation. A valid instrument will support a higher quality learning process by providing accurate data on students' competency achievement, which can be used as evaluation material and for continuous learning improvement.

Results of the Construct Validity of the HOTS Instrument

After conducting content validity testing of the HOTS instrument with experts, construct validity testing was also carried out on students in the field to assess the validity or legitimacy of the instrument to be used (Siallagan et al., 2023). The validity test used was construct validity. Construct validity in this study was conducted to evaluate the extent to which the

items in the HOTS instrument consistently and representatively measure higher-order thinking skills in early childhood according to the theoretical construct designed. Construct validity focuses on the correlation between each item and the overall construct represented by the total score. Using the SPSS program and involving 30 children aged 4–6 years as respondents, the table value of r was obtained as 0.361 (with df = 28 and α = 0.05). Based on general criteria, an item is considered valid if the calculated r value is greater than the table r value (Khoirunnisa & Vinayastri, 2021).

Table 5 Results of Item Validity Test on Students of Group B, Al Quran Al Hakim Kindergarten

No	Student Initials	Test Item Number															Total
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	AP	1	0	1	1	1	0	1	0	0	0	0	0	1	1	1	8
2	ANA	1	0	0	1	1	1	0	1	0	1	1	1	0	0	1	9
3	ACM	1	0	1	1	0	0	0	0	0	0	0	0	1	0	1	5
4	AT	0	1	0	0	0	1	0	1	1	1	0	1	0	0	0	6
5	AFA	1	0	1	1	0	0	0	0	0	0	0	0	1	0	1	5
6	DF	0	1	1	0	1	1	1	1	1	0	1	1	1	1	0	11
7	DMS	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
8	DAA	1	0	1	1	1	0	1	0	1	0	1	0	1	1	1	10
9	DPP	0	1	0	0	0	1	0	1	0	1	0	1	0	0	0	5
10	DH	1	0	1	1	1	0	1	0	0	1	1	0	1	1	1	10
11	ER	0	1	0	0	0	1	0	1	1	0	0	1	0	0	0	5
12	FAP	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	3
13	FNA	0	0	1	0	1	0	1	0	1	0	1	0	1	1	0	7
14	FA	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	3
15	FAA	1	0	1	1	0	0	1	0	1	1	1	0	1	1	1	10
16	FP	1	1	0	1	1	1	0	1	0	1	0	1	0	0	1	9
17	GF	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	12
18	HZTF	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	3
19	HAD	1	1	1	1	0	0	1	1	0	1	0	1	1	1	1	11
20	ISL	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	3
21	IA	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	4
22	JR	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	11
23	JNA	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	5
24	KHS	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	12
25	MA	1	1	0	1	0	1	0	1	0	1	0	1	0	0	1	8
26	MNR	1	0	1	1	1	0	1	0	1	1	1	1	1	1	1	12
27	NNA	0	0	1	1	1	1	0	0	1	0	0	0	1	0	0	6
28	NA	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	13
29	PZT	1	0	0	1	1	1	0	0	0	1	0	0	0	0	1	6
30	QNI	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	8

Analysis of Valid Test Items (10 items), items that have a calculated r value > 0.361 and are declared valid are: Item 1 (r = 0.480), indicating a fairly strong correlation with the total score, meaning that this item is capable of measuring the relevant HOTS dimension. Most likely, this item involves the ability to group or compare simple objects that are familiar to the child. Item 2 (r = 0.467), relatively high validity indicates that this item is effective in triggering analytical thinking processes or simple logical associations, for example connecting cause and effect. Item 3 (r = 0.608), high correlation shows that this item can encourage the child to perform synthesis or identify patterns of activity that reflect higher order thinking ability. Item 4 (r = 0.585), this item appears to involve problem-solving in a concrete context, for example choosing strategies to solve simple challenges appropriate to the child's age. Item 7 (r = 0.711), one of the items with the highest correlation, indicates that this question may encourage the child to make decisions based on logical thinking or consider alternative solutions.

Item 8 (r = 0.495), valid and fairly strong, this item may trigger predictive or exploratory responses from the child, for example predicting the outcome of an action. Item 12 (r = 0.587), high validity indicates the ability of this item to stimulate idea elaboration or causal explanation from early childhood learners. Item 13 (r = 0.509), this item appears to require the child to classify or organize ideas based on logical categories, which is part of the analytical process. Item 14 (r = 0.808), this is the test item with the highest correlation, indicating that this item is highly representative of the overall HOTS construct. It is strongly suspected that this item involves divergent thinking activities (for example creating something from existing materials or composing a story from images). Item 15 (r = 0.674), highly valid, indicates that the child can respond meaningfully to stimuli that encourage evaluative or reflective abilities in a concrete form.

These results show that most of the valid items are closely related to higher-order thinking activities typical of early childhood, such as grouping, predicting, analyzing cause-and-effect relationships, choosing strategies, and simple creative actions. These items appear to have been designed by taking into account concrete contexts, simple language, and visual/verbal stimuli that align with the developmental characteristics of children aged 4–6 years. This is in line with the opinion of (Purnama et al., 2021), which states that assessment for early childhood should prioritize meaningful activity-based contexts, visual elements, and direct engagement.

Analysis of Invalid Test Items (5 items), five test items with calculated r values < 0.361 are declared invalid, namely: Item 5 (r = 0.033), the correlation value is very low, almost approaching zero. This indicates that the item has almost no correlation with the total score. It is strongly suspected that the item is too abstract or uses language/instructions that are not understood by early childhood learners. Item 6 (r = 0.116), the low r value indicates that responses to this item are highly variable and do not align with the expected construct pattern. It is possible that the item contains ambiguous or insufficiently contextual stimuli. Item 9 (r = 0.083), similar to items 5 and 6, this item has a low correlation, indicating that children were unable to understand the item or that the question did not require higher-order thinking activities. Item 10 (r = 0.298), although approaching the critical value, this item still does not meet validity criteria. This suggests that the item construction may include non-functioning distractors or confusing instructions. Item 11 (r = 0.131), the low correlation value indicates that this item may be too easy or, conversely, too difficult, thus failing to differentiate HOTS abilities among individuals. According to (Apriyansyah et al., 2023), invalid test items in early childhood assessments are generally caused by sentence formulations that do not align with children's linguistic patterns, the use of non-communicative images, or instructions that are too complex. This condition further supports the need for modification and alignment of visual and verbal stimuli in accordance with the cognitive development characteristics of children.

In general, the failure of validity in these five items may be caused by several factors, including: (1) the mismatch of item wording with the language level of early childhood learners; (2) unclear or unappealing stimuli or illustrations; (3) cognitive indicators that are either too high or too low; and (4) a lack of everyday context that prevents children from connecting with the item. Therefore, revisions of these items are strongly recommended, either through language simplification, improvement of images/visuals, or the contextual shift to play-based or everyday life activities of the children.

Table 6 Construct Validity Test of HOTS Instrument

No.	r calculated	r table	Description
1	0,480	0,361	Valid
2	0,467	0,361	Valid
3	0,608	0,361	Valid
4	0,585	0,361	Valid
5	0,033	0,361	Not Valid
6	0,116	0,361	Not Valid
7	0,711	0,361	Valid
8	0,495	0,361	Valid
9	0,083	0,361	Not Valid
10	0,298	0,361	Not Valid
11	0,131	0,361	Not Valid
12	0,587	0,361	Valid
13	0,509	0,361	Valid
14	0,808	0,361	Valid
15	0,674	0,361	Valid

Preliminary Conclusion of Construct Validity Test Results, namely from the 15 HOTS instrument items developed, 10 items (66.7%) were declared valid, while 5 items (33.3%) were not valid. This percentage indicates that the majority of the instrument items possess good construct validity quality and can represent higher-order thinking abilities in early childhood within a concrete context. This result serves as an important basis for filtering test items in the next phase, namely the reliability test, while also providing material for reflection in the revision and development of the instrument during a broader implementation phase.

The next step after obtaining the construct validity test results is the selection of the 10 items that have been declared valid as part of the Higher Order Thinking Skills (HOTS) instrument for early childhood. These ten items will be used as the main instrument to be further examined in the reliability testing phase. This process aims to determine the extent to which the selected items demonstrate good internal consistency, i.e., whether the items can yield stable and consistent measurement results when used in different situations or with different groups of children.

The reliability test will be conducted using the SPSS program through the calculation technique of Cronbach's Alpha, which is appropriate for measuring the reliability of non-dichotomous instruments used in early childhood assessment. Through this test, a reliability coefficient value will be obtained, which describes the level of consistency among the test items in measuring the HOTS construct. If the obtained reliability value meets the minimum standard (e.g., ≥ 0.60), then the instrument can be considered feasible for use in assessing higher-order thinking skills in early childhood. This process will simultaneously strengthen the overall validity of the instrument by integrating both validity and reliability aspects as essential quality indicators of a measurement tool.

Results of the HOTS Instrument Reliability Test Using Cronbach's Alpha

Reliability is a fundamental dimension in evaluating the quality of psychometric instruments, reflecting the consistency, stability, and replicability of measurement results for a psychological construct (Subhaktiyasa, 2024). In this context, the intended construct is Higher Order Thinking Skills (HOTS), namely high-level cognitive abilities that include analyzing, evaluating, and creating, as formulated in the revised Bloom's Taxonomy (Anderson & Krathwohl, 2001) in (Marta et al., 2025).

The reliability test using the Cronbach's Alpha method was chosen because the instrument consists of several items with ordinal-interval scales, and alpha allows for the measurement of internal consistency coefficients without requiring retesting (test-retest). Philosophically, this approach assumes that each item in the instrument is a reflective indicator of the same latent construct, namely HOTS in early childhood.

The instrument tested consists of 10 selected items that have passed the construct validation, namely item numbers 1, 2, 3, 4, 7, 8, 12, 13, 14, and 15. Data were obtained from 30 early childhood respondents, representing the target population of the HOTS assessment in early childhood education units. This number meets the minimum threshold for exploratory reliability testing, which is 10–30 subjects (Budiastuti & Bandur, 2018), with no missing data, thus allowing for a complete and representative analysis.

Table 7 Interpretation of Cronbach's Alpha Coefficient

Case Processing Summary			
		N	%
Cases	Valid	30	100.0
	Excluded[a]	0	.0
	Total	30	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics	
Cronbach's Alpha	N of Items
.838	10

Interpretation of cronbach's alpha coefficient: high reliability indicator. The test result shows a Cronbach's Alpha value of 0.838, which indicates a high level of internal reliability (good), according to the classification by George & Mallery (2003) in (Saidi & Siew, 2019) and is also consistent with the interpretation of (Rapareni et al., 2024) which states that Alpha ≥ 0.60 is appropriate for measuring psychological constructs in social and educational studies. This value indicates that the items in the instrument exhibit strong logical and empirical interrelations, thus the measurement results are considered stable and minimally affected by random error. In early childhood assessment, high reliability is crucial given the high likelihood of response fluctuations due to the still very dynamic psychological development of young children.

Tabel 8 Item Statistical Distribution

Item Statistics			
	Mean	Std. Deviation	N
Soal_1	.6000	.49827	30
Soal_2	.4000	.49827	30
Soal_3	.5667	.50401	30
Soal_4	.6000	.49827	30
Soal_7	.4667	.50742	30
Soal_8	.4333	.50401	30
Soal_12	.4667	.50742	30
Soal_13	.5333	.50742	30
Soal_14	.4333	.50401	30
Soal_15	.5333	.50742	30

Item statistical distribution: indication of balance and sensitivity. The results of the descriptive statistical analysis indicate that the average item scores range between 0.40 and 0.60, with standard deviations between 0.49 and 0.50. This range suggests that the test items are at a moderate level of difficulty and are capable of proportionally distinguishing children's abilities. The relatively symmetrical distribution indicates that there are no items that are too easy or too difficult, which, if present, could compromise the accuracy and reliability of the measurement. Thus, this instrument has demonstrated good sensitivity to variations in Higher Order Thinking Skills (HOTS) among early childhood learners.

Table 9 Item Total Correlation

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Item_1	4.4333	8.737	.450	.830
Item_2	4.6333	8.930	.380	.837
Item_3	4.4667	8.257	.621	.814
Item_4	4.4333	8.461	.552	.821
Item_7	4.5667	8.254	.617	.814
Item_8	4.6000	8.938	.371	.838
Item_12	4.5667	8.737	.438	.832
Item_13	4.5000	8.534	.512	.825
Item_14	4.6000	7.972	.732	.803
Item_15	4.5000	8.190	.641	.812

Item-total correlation: evaluating item contribution to reliability. The corrected correlation between each item and the total score (Corrected Item-Total Correlation) is a key indicator for evaluating how strongly an item reflects the overall construct. The results show that Item 14 (r = 0.732) and Item 15 (r = 0.641) have the highest correlations, indicating that these items are the most representative in capturing the overall HOTS abilities. Item 2 (r = 0.380) and Item 8 (r = 0.371) show the lowest correlations, yet they remain above the minimum threshold of 0.30 established in classical psychometrics. No items fall below the 0.30 cutoff, indicating that there is no need for item elimination or revision. This implies that each item contributes meaningfully to the construct, and internal consistency is fairly evenly distributed.

Tabel 10 Item Deletion Analysis

Scale Statistics			
Mean	**Variance**	**Std. Deviation**	**N of Items**
5.0333	10.309	3.21079	10

Item deletion analysis: reliability resilience to modification. The column “Cronbach’s Alpha if Item Deleted” indicates that the removal of any single item would not significantly increase the Alpha value, and in some cases, would actually reduce the total Alpha. In fact, the highest Alpha obtained after item deletion (0.838) is identical to the overall value, meaning that no single item significantly disrupts the instrument’s reliability. This interpretation demonstrates that the structure of the instrument is complementary and mutually reinforcing across items, which is a critical requirement for developing a standardized and balanced HOTS measurement scale.

Critical reflection and implications for hots assessment in early childhood. Based on the high reliability results, several important implications can be drawn: The assessment of Higher Order Thinking Skills (HOTS) in early childhood is highly feasible when carried out systematically using this instrument as a reliable measurement tool. The instrument can be utilized in the context of formative assessment, diagnostic purposes, or quasi-experimental research to measure changes in higher-order thinking skills. Within the framework of early childhood education (ECE), this high reliability provides justification that HOTS is not an exclusive domain of higher educational levels, but can be developed from an early age, provided that the instrument is aligned with the child's cognitive developmental stage.

Table 11 Instrument Reliability Results

Reliability Statistics	
Cronbach's Alpha	**N of Items**
,838	10

Based on the results of the reliability analysis using Cronbach’s Alpha, which yielded a value of 0.838, it can be concluded that: This instrument for measuring HOTS in early childhood meets the criteria for high internal consistency, with no problematic items identified. All items demonstrate adequate correlation with the total score, and there is no redundancy or need for item elimination. Overall, this instrument is appropriate for use as a valid and reliable assessment tool, with considerable potential for further development through advanced testing (such as factor analysis or inter-rater reliability testing if applied in observational settings).

CONCLUSION

Based on the results of the validity and reliability analyses, it can be concluded that the developed HOTS instrument is feasible and meets the eligibility standards for use as a measurement tool for students’ Higher Order Thinking Skills (HOTS). First, in terms of content validity, the instrument received an average total score of 86.1% from three subject matter experts, which, according to Arikunto’s criteria (2005), falls into the "good" and valid category. The assessment covered three main aspects: content, language, and presentation, all of which indicated the instrument’s adequate quality. Second, the results of the construct validity test on 15 items using the SPSS program with a sample of 30 students showed that 10 items were declared valid, as the calculated r values were greater than the critical r value (0.361). This indicates that most items successfully measured the intended construct, although several items need to be revised or eliminated. Third, the reliability test results for the 10 validated items using the Cronbach’s Alpha method yielded a coefficient of 0.838, which is well above the minimum threshold of 0.60. This demonstrates that the instrument possesses high internal consistency and can be relied upon if used repeatedly within the same measurement context. Thus, the 10 HOTS instrument items (items 1, 2, 3, 4, 7, 8, 12, 13, 14, and 15) are declared both valid and reliable, and can be used as an assessment tool to measure students’ higher order thinking skills in learning contexts.

REFERENCES

1. Anshari, M. I., Nasution, R., Irsyad, M., Alifa, A. Z., & Zuhriyah, I. A. (2024). Validity and Reliability Analysis of Summative Final Semester Odd Items for Islamic Education Subject. Edukatif: Journal of Educational Science, 6(1), 964–975. https://doi.org/10.31004/edukatif.v6i1.5931
2. Apriyansyah, C., Tjalla, A., Saptono, A., & Sukatmi, S. (2023). The Importance of Modifying Instruments for Holistic-Integrative Early Childhood Development. Obsesi: Journal of Early Childhood Education, 7(6), 6991–7002. https://doi.org/https://doi.org/10.31004/obsesi.v7i6.5338

3. Arikunto, S. (2008). *Fundamentals of Educational Evaluation.* Bumi Aksara.
4. Azwar, S. (2017). *Reliability and Validity* (4th ed.). Pustaka Pelajar.
5. Budiastuti, D., & Bandur, A. (2018). *Research Validity and Reliability: Complete with Analysis Using NVIVO, SPSS, and AMOS. Mitra Wacana Media.*
6. Christianti, M. (2024). Validity and Reliability of the Early Literacy Assessment (ALIA) for Early Childhood. Journal of Child Education, 13(2), 250–264. https://doi.org/https://doi.org/10.21831/jpa.v13i2.649
7. Faizah, U. (2019). Authentic Assessment Model to Evaluate Character Achievement in Kindergarten / Raudlatul Athfal. MUKADDIMAH: Journal of Islamic Studies, 4(1), 1–37
8. Fitriani, S. S. A., & Vinayastri, A. (2022). Development of Critical Thinking Ability Instrument for Early Childhood. Pedagogy: Journal of Early Childhood and Early Childhood Education, 8(1), 21. https://doi.org/10.30651/pedagogi.v8i1.8973
9. Frausel, R., Silvey, C., Freeman, C., Dowling, N., & Goldin-Meadow, S. (2020). The origins of higher-order thinking lie in children's spontaneous talk across the pre-school years. *Cognition*, *200*. https://doi.org/https://doi.org/10.1016/j.cognition.2020.104274.
10. Kaaffah, R. R. S., Wijiyono, A. W., & Rahmayanti, I. (2021). Content Validity of the Evaluation Tools for Indonesian Language Textbooks for 10th Grade Senior High School Students. Imajeri: Journal of Indonesian Language and Literature Education, 3(2), 158–167. https://doi.org/https://doi.org/10.22236/imajeri.v3i2.6572
11. Khoirunnisa, M. F., & Vinayastri, A. (2021). Development of Fine Motor Skills Instrument for Early Childhood. Golden Age Journal, Hamzanwadi University, *5*(02), 356–365.
12. Kunduracı, H., Yaralı, K., & Kaynak, S. (2024). Measuring parental behaviors supporting higher order thinking skills in children: A scale development study. *Thinking Skills and Creativity.* https://doi.org/https://doi.org/10.1016/j.tsc.2024.101685.
13. Li, W., Huang, J., Liu, C., Tseng, J., & Wang, S. (2023). A study on the relationship between student' learning engagements and higher-order thinking skills in programming learning. *Thinking Skills and Creativity.* https://doi.org/https://doi.org/10.1016/j.tsc.2023.101369.
14. Marada, R., Nusantari, E., & Dama, L. (2021). Development of a Higher Order Thinking Skills (HOTS)-Based Instrument to Train Students' Critical Thinking Skills in Biology Subjects. *Jurnal Normalita*, *9*(2), 188–194. https://doi.org/https://ejurnal.pps.ung.ac.id/index.php/JN/article/view/441
15. Marta, M. A., Purnomo, D., & Gusmameli, G. (2025). The Concept of Bloom's Taxonomy in Learning Design. *Lencana: Journal of Innovation in Educational Sciences*, *3*(1). https://doi.org/https://doi.org/10.55606/lencana.v3i1.4572
16. Nisa, N. C., Nadiroh, N., & Siswono, E. (2018 Higher Order Thinking Skills (HOTS) about the Environment Based on Students' Academic Backgrounds. *PLPB: Journal of Environment Education and Sustainable Development, 19*(02), 1–14. https://doi.org/https://doi.org/10.21009/PLPB.192.01
17. Paudpedia. (2022). *Development of Higher Order Thinking Skills (HOTS)* in Early Childhood Education. Directorate of Early Childhood Education, Directorate General of Early Childhood Education, Primary Education, and Secondary Education, Ministry of Education, Culture, Research, and Technology.
18. Piaget, J. (1952). *The Origins of Intelligence in Children*. International Universities Press.
19. Purnama, S., Ulfah, M., Susilo, E., Mutmainnah, M., & Amalia, R. (2021). *Assessment of Early Childhood Development (M. A. Latif, Ed.; 1st ed.).* CV Multiartha Jatmika Yogyakarta.
20. Purnasari, P. D., Silvester, S., & Lumbantobing, W. L. (2021). Development of Higher Order Thinking Skills (HOTS) Assessment Instrument Viewed from Students' Learning Styles. *Sebatik*, 25(2), 571–580. https://doi.org/10.46984/sebatik.v25i2.1607
21. Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. *Journal on Education*, *6*(2), 10967–10975. https://doi.org/10.31004/joe.v6i2.4885
22. Rapareni, Y., Yulanda, D., & Oktala, R. (2024). The Effect of Locus of Control and Parental Support on the Quality of Graduates from Serelo University Lahat. *Trending: Journal of Economics, Accounting, and Management*, 2(3).
23. Rodiana, S., & Pahlevi, T. (2020). Development of a Higher Order Thinking Skills (HOTS)-Based Assessment Instrument in Archival Subjects for the Office Administration Program at SMKN 1 Sooko Mojokerto. *Journal of Office Administration Education (JPAP), 8(1),* 82–95. https://doi.org/10.26740/jpap.v8n1.p82-95
24. Safi'i, I., Tarmini, W., & Yanti, P. G. (2022). Implementation of Creative-Innovative Aspects in the Indonesian Language BSE Evaluation Instrument. *RETORIKA: Jurnal Bahasa, Sastra, Dan Pengajarannya*, *15*(1), 52–58. https://doi.org/10.26858/retorika.v15i1.17546
25. Saidi, S. S., & Siew, N. M. (2019). Investigating the Validity and Reliability of Survey Attitude towards Statistics Instrument among Rural Secondary School Students. *International Journal of Educational Methodology*, *5*(4), 651–661. https://doi.org/10.12973/ijem.5.4.651
26. Setiawati, W., Asmira, O., Ariyana, Y., Bestary, R., & Pudjiastuti, A. (2019). Assessment Book Oriented Towards Higher Order Thinking Skills. Directorate General of Teachers and Education Personnel, Ministry of Education and Culture.

27. Siallagan, F., Tambunan, L. O., & Sidabutar, R. (2023). Development of Higher Order Thinking Skill (HOTS) Test Instrument on Number Material for Grade VII Students at SMP Negeri 2 Siantar. INNOVATIVE: Journal of Social Science Research, 3(6), 8990–9004.
28. Subhaktiyasa, P. G. (2024). Evaluation of Validity and Reliability of Quantitative Research Instruments: A Literature Study. *Journal of Education Research*, *5*(4), 5599–5609. https://doi.org/https://doi.org/10.37985/jer.v5i4.1747
29. Sulaiman, S. (2020). Higher order thinking skill (Hots) Pada Anak Usia Dini. *SULOH: Jurnal Bimbingan Konseling Universitas Syiah Kuala*, *5*(1), 1–10. https://doi.org/https://doi.org/10.24815/suloh.v5i2.17732
30. Utsman, U. (2013). Development of Assessment Instruments for Early Childhood. Universitas Negeri Yogyakarta.
31. Wantah, M. J. (2010). Validity and Reliability of Assessment Instruments for Kindergarten Children with Special Needs. *Jurnal Penelitian Dan Evaluasi Pendidikan*, https://doi.org/10.21831/pep.v14i1.1979
32. Wewe, M., & Wangge, M. C. T. (2021). Development of a High Order Thinking Skill (HOTS)-Based Test Instrument on the Material of Two-Variable Linear Equation Systems. *Jurnal Pendidikan Tambusai*, *5*(3), 10693–10702. https://doi.org/https://doi.org/10.31004/jptam.v5i3.2690
33. Yıldız, C., & Yildiz, T. (2021). Exploring the relationship between creative thinking and scientific process skills of preschool children. *Thinking Skills and Creativity.* https://doi.org/https://doi.org/10.1016/J.TSC.2021.100795.
34. Zebua, N. (2024). Literature Study: The Role of Higher Order Thinking Skills in the Learning Process. *Edukasi Elita: Journal of Educational Innovation*, *1*(2), 92–100. https://doi.org/https://doi.org/10.62383/edukasi.v1i2.110