 ISSN 3025-2644

Validity and reliability of a future-oriented critical thinking skills diagnostic test in elementary science education

Efi Kurniasari [a,b1,*], Sugiyarto Surono [b,2], Ishafit [b,3], Selviana Veronika Moruk [c,4]

[a] Universitas Muhammadiyah Sukabumi, Sukabumi, Indonesia
[b] Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[c] Universidade Oriental Timor Lorosae, Timor-Leste
[1] efikurniasari@ummi.ac.id*; [2] sugiyarto@math.uad.ac.id; [3] ishafit@pfis.uad.ac.id; [4] morukselviana1@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received: May 4, 2026
Revised: June 16, 2026
Accepted: June 17, 2026

Keywords
Critical thinking
Diagnostic test
Elementary school
FOCTS
Validity and reliability

ABSTRACT

Future-Oriented Critical Thinking Skills (FOCTS) extend conventional critical thinking by incorporating a future-oriented perspective that enables students to generate innovative solutions to emerging challenges. However, no validated diagnostic instrument currently exists to measure this construct among elementary school students. This study aimed to develop and evaluate the validity and reliability of a FOCTS diagnostic test for integrated science and social studies. The novelty of this study lies in operationalizing the future-oriented dimension of critical thinking into a measurable diagnostic instrument supported by both logical and face validity evidence. The subjects used involved 30 students from one elementary school. Content validity was assessed through expert judgment covering logical validity (construct alignment) and face validity (clarity, readability, and presentation), using Aiken's V coefficient. Empirical validity was examined using Pearson's product-moment correlation, while reliability was evaluated using Cronbach's Alpha. Due to scheduling constraints, empirical testing involved 30 sixth-grade students from one elementary school. The results showed an overall Aiken's V value of 0.74, indicating adequate content validity. Empirical testing revealed that 25 of the 30 items were valid r = 0.367–0.713; r-critical = 0.361, α = 0.05, while five items were invalid. The 25 valid items demonstrated high reliability Cronbach's Alpha = 0.871. These findings indicate that the FOCTS instrument has promising psychometric properties, although further validation with larger and more representative samples is recommended before wider implementation. The findings imply that the FOCTS instrument can serve as a promising diagnostic tool for identifying elementary students' future-oriented critical thinking strengths and needs, pending further validation with larger samples.

1. Introduction

Critical thinking has long been positioned as a core competency for 21st-century learners, encompassing skills such as analysis, interpretation, evaluation, and inference applied to existing problems and information (Dilekçi & Karatay, 2023; Ennis, 2018). It is widely regarded as one of the highest educational priorities for modern societies, closely linked to problem-solving, adaptability, and academic achievement (Schäfer et al., 2024), and bibliometric reviews indicate a steadily growing body of research on critical thinking instruction at the primary and secondary levels over the past decade (Loyens et al., 2023). However, contemporary educational discourse increasingly recognizes that conventional critical thinking frameworks, while valuable, are primarily retrospective and analytical in nature, focused on evaluating present or past information rather than generating novel solutions for problems that have not yet emerged.

This limitation is particularly salient in the era of rapid technological and societal change, where students must not only analyze existing conditions but also anticipate, innovate, and design solutions for future challenges (González-Pérez & Ramírez-Montoya, 2022; Thornhill-Miller et al., 2023). Systematic reviews of critical thinking instruction at early educational levels have noted that the field still lacks well-validated, developmentally appropriate assessment tools, particularly those capturing forward-looking or generative dimensions of thinking. In the Indonesian elementary education context, the integration of critical thinking into integrated science and social studies has become increasingly emphasized, which prioritizes eight dimensions of graduate profiles, including critical thinking (Aristanti & Fatayan, 2024; Musfidah et al., 2023).

To address this conceptual gap, the construct of FOCTS was developed as a theoretical extension of critical thinking. FOCTS explicitly incorporates a generative, anticipatory dimension, future innovation, positioned at Bloom's Revised Taxonomy level C6 (Create), reflecting the capacity to design novel ideas and solutions grounded in scientific understanding (Suhendar et al., 2025). The FOCTS construct consists of five interrelated indicators: Problem Understanding, Analysis, Evaluation, Inference, and Future Innovation (Alshaikh, 2024). While the first four indicators parallel components found in established critical thinking models, the explicit integration of a future-oriented, generative component distinguishes FOCTS as a theoretically distinct construct rather than a relabeling of existing frameworks. This distinction is conceptually significant because it shifts the assessment focus from merely diagnosing students' analytical capacity toward capturing their readiness to engage with open-ended (Sarwanto et al., 2020), future-relevant problem solving, a dimension increasingly demanded by 21st-century educational goals but rarely operationalized in existing instruments (Care et al., 2020).

The relevance of FOCTS is further situated within the Game Solve Learning Model, an instructional model that integrates problem-based learning with game-based elements to actively engage elementary students in scientific inquiry (Kurniasari et al., 2026). Within this model, students are confronted with structured problem scenarios embedded in game mechanics, requiring them to understand the problem, analyze relevant information, evaluate possible solutions, draw logical inferences, and ultimately propose innovative resolutions a sequence that directly mirrors the five FOCTS indicators. The six science topics selected for instrument development (vision and light, hearing and sound, ecosystem harmony, magnets, electricity and technology, and Earth science) represent core fifth-grade integrated science and social studies content areas in which Game Solve learning activities are typically implemented, ensuring that the diagnostic instrument is contextually aligned with the pedagogical model it is intended to evaluate. Assessing the effectiveness of Game Solve learning in fostering FOCTS, however, requires an instrument with demonstrated validity and reliability; without such evidence, conclusions drawn from the instrument's scores would lack scientific credibility.

Validity and reliability are foundational requirements for any educational measurement instrument. Validity refers to the degree to which an instrument measures the construct it is intended to measure, while reliability refers to the consistency of measurement results across repeated administrations (Clark & Watson, 2019; Ramadhan et al., 2024). In the context of newly developed instruments, content validity, established through expert judgment, is typically examined first, followed by empirical validity and reliability testing using actual respondent data. Content validity itself encompasses two complementary forms: logical validity, which concerns the theoretical and conceptual alignment between items and the construct they are intended to represent, and face validity, which concerns the clarity, readability, and technical adequacy of items as judged by experts and, ideally, by representative respondents (Aini & Putri, 2024; Almanasreh et al., 2019). To date, no published diagnostic instrument has examined both the logical and face validity dimensions of a FOCTS-based construct for elementary school students, nor has the empirical validity and reliability of such an instrument been established. This gap limits the ability of teachers and researchers to objectively measure students' FOCTS.

Based on this background, this study aimed to develop and validate a diagnostic test instrument for measuring FOCTS in integrated science and social studies learning. The novelty of this study is the focus on assessing future-oriented critical thinking, which has received limited attention in elementary education. The instrument was evaluated through content validity, empirical validity, and reliability testing using Aiken's V, Pearson's product-moment correlation, and

Cronbach's Alpha. The results provide evidence that the instrument can be used for research and classroom assessment. Specifically, this study sought to answer the following research questions: To what extent does the FOCTS diagnostic test instrument demonstrate adequate content validity, as evidenced by both logical validity and face validity through expert judgment using Aiken's V coefficient? To what extent are the items of the FOCTS diagnostic test instrument empirically valid, as evidenced by Pearson's product-moment correlation between item scores and total test scores? To what extent does the FOCTS diagnostic test instrument demonstrate adequate internal consistency reliability, as evidenced by Cronbach's Alpha coefficient?

2. Method

2.1. Research design

This study employs a research and development approach to design a diagnostic test instrument to measure FOCTS and to test its content validity, empirical validity, and reliability. Instrument development follows a sequential procedure beginning with the formulation of a theoretical framework, definition of constructs and indicators, the development of an item matrix, item writing, expert validation, and empirical testing. This study focuses on establishing the basic psychometric evidence of content validity, empirical validity, and reliability required before an instrument can be considered for wider use.

2.2. Research subjects

The content validity phase involved three expert validators representing the fields of educational evaluation, science education, and classroom teaching practices in elementary schools. Their feedback was used to evaluate the relevance, clarity, and representativeness of each instrument item before empirical testing. For the empirical testing phase, the instrument was administered to 30 students from one elementary school to examine its psychometric properties and overall feasibility.

2.3. Research instrument

The instrument developed in this study was a diagnostic test consisting of 30 multiple-choice items with four answer options (A, B, C, D). Each item was constructed to represent one of five FOCTS indicators: Problem Understanding, Analysis, Evaluation, Inference, and Future Innovation, distributed across integrated science and social studies topics, including vision and light, hearing and sound, ecosystem harmony, magnets, electricity and technology, and Earth science. The items were designed to assess students' higher-order thinking skills within meaningful scientific contexts while ensuring balanced representation of each indicator and content domain. The instrument blueprint was developed to align the assessment indicators with the targeted learning topics, thereby supporting the validity of content coverage. The item indicators and corresponding topics are presented in Table 1.

2.4. Research procedure

Instrument development was conducted in four stages. First, in the design stage, the FOCTS theoretical framework was operationalized into constructs, indicators, an item matrix, and 30 draft multiple-choice items (Badenes-Ribera et al., 2020). This stage ensured that each item was systematically aligned with the intended FOCTS indicators and the targeted integrated science and social studies learning content. Second, during the expert validation stage, three expert validators assessed each item using a four-point rating scale across three aspects: item construction, substance/content, and format and technical quality. Third, in the empirical testing stage, the revised instrument with 30 items was administered to 30 students described in the Research Subjects section. Fourth, in the data analysis stage, item responses were analyzed for empirical validity and reliability, as described in the following section.

2.5. Data analysis

Content validity data from the three expert validators were analyzed using Aiken's V coefficient for each item, aspect, and indicator. Empirical validity was analyzed using Pearson's product-moment correlation between the score of each item and the total test score, calculated using the Statistical Package for the Social Sciences. The reliability of items that were found to be valid was then analyzed using Cronbach's Alpha. The expert judgment and empirical testing confirmed the quality of the developed instrument. A summary of the analysis criteria applied in the content validity, empirical validity, and reliability stages is presented in Table 2.

Table 1. Item indicators and topics of the FOCTS instrument

No	FOCTS indicator	Taxonomy level	Material	Question indicator
1	Problem Understanding	C4	Seeing because of light	Students are able to analyze the role of light in the process of seeing.
2		C4	Hearing because of sound	Students are able to analyze the sound process until it can be heard
3		C4	Harmony in the ecosystem	Students are able to understand the relationship between living things in the ecosystem
4		C4	Magnet	Students are able to identify the properties of objects that can be attracted by magnets.
5		C4	Electricity and technology	Students are able to understand the role of electricity in everyday life
6		C4	Our Earth	Students are able to analyze the basic conditions of the earth
7	Analysis	C4	Seeing because of light	Students are able to describe the relationship between light and sight.
8		C4	Hearing because of sound	Students are able to describe the sound propagation process
9		C4	Harmony in the ecosystem	Students are able to analyze cause-effect relationships in ecosystems
10		C4	Magnet	Students are able to compare magnetic and non-magnetic objects
11		C4	Electricity and technology	Students are able to analyze simple electrical circuits
12		C4	Our Earth	Students are able to analyze natural phenomena that occur on earth
13	Evaluation	C5	Seeing because of light	Students are able to assess conditions that affect vision
14		C5	Hearing because of sound	Students are able to assess the impact of sound on hearing
15		C5	Harmony in the ecosystem	Students are able to assess human actions towards ecosystems
16		C5	Magnet	Students are able to assess the use of magnets in life
17		C5	Electricity and technology	Students are able to assess the use of electricity wisely
18		C5	Our Earth	Students are able to assess efforts to protect the earth
19	Inference	C5	Seeing because of light	Students are able to conclude the role of light in life
20		C5	Hearing because of sound	Students are able to conclude the process of sound production
21		C5	Harmony in the ecosystem	Students are able to conclude the importance of ecosystem balance
22		C5	Magnet	Students are able to conclude the properties of magnets
23		C5	Electricity and technology	Students are able to conclude the benefits of electricity
24		C5	Our Earth	Students are able to conclude how to protect the earth
25	Future Innovation	C6	Seeing because of light	Students are able to design the use of light in life
26		C6	Hearing because of sound	Students are able to create simple sound-based tools
27		C6	Harmony in the ecosystem	Students are able to design solutions to maintain ecosystem balance.
28		C6	Magnet	Students are able to develop ideas for using magnets
29		C6	Electricity and technology	Students are able to design electricity saving solutions
30		C6	Our Earth	Students can make innovations to protect the earth

Table 2. Summary of item analysis criteria

Aspect	Criteria
Item Validity	r-calculated ≥ 0.361 → Valid
Instrument Reliability	Cronbach's Alpha > 0.70 → Reliable

3. Results and Discussion

Prior to empirical testing, the 30-item instrument was evaluated by three expert validators using Aiken's V coefficient to determine its content validity. This evaluation examined the relevance and appropriateness of each item based on the established assessment criteria and FOCTS indicators. The results of the content validity analysis are presented in Table 3.

Table 3. Results of content validity analysis using Aiken's V

No	Aspect	Indicator	V1	V2	V3	Σs	V	Mean V
1	Item Construction	Each item has one clear competency focus	4	4	4	9	0.75	
		Item statement is unambiguous and easy to understand	4	4	4	9	0.75	
		Answer alternatives are homogeneous, logical, and contextually appropriate	4	4	4	9	0.75	0.73
		Sentence length is appropriate for students' cognitive developmental level	4	3	4	8	0.67	
2	Substance/ Content	Indicators measure FOCTS aspects comprehensively	4	4	4	9	0.75	
		Aspects measured cover all theoretical foundations of the learning design	4	4	3	8	0.67	
		Each item aligns with the FOCTS measurement objectives	4	4	4	9	0.75	0.73
		Item wording is clear, specific, and does not allow multiple interpretations	4	4	4	9	0.75	
3	Format & Technical	Writing uses correct Bahasa Indonesia	4	4	4	9	0.75	
		Layout is neat, consistent, and easy to read	3	3	4	7	0.78	0.76
		Test instructions are clear, concise, and easy to understand	4	4	4	9	0.75	
		Overall Mean Aiken's V						0.74

The content validity analysis results show that all three aspects achieved Aiken's V values at or above the acceptable threshold of 0.70: Item Construction V = 0.73, Substance/Content V = 0.73, and Format and Technical Quality V = 0.76, yielding an overall mean Aiken's V = 0.74. V value of 0.70 or above indicates adequate content validity based on expert judgment. The involvement of three validators from distinct professional backgrounds educational evaluation, science education, and classroom practice strengthens the credibility of this assessment, as it reflects a multidimensional review encompassing the instrument's theoretical alignment, linguistic clarity, and technical quality. The high ratings for the Substance/Content aspect confirm that the five FOCTS indicators are theoretically grounded and meaningfully operationalized within the multiple-choice format. This is particularly important for the Future Innovation indicator (C6), which represents the highest cognitive demand and is most susceptible to content validity concerns in objective test formats (Liu et al., 2024; Morad et al., 2021).

Two sub-indicators recorded V = 0.67, slightly below the 0.70 threshold specifically 'sentence length appropriateness for students' cognitive level' and 'theoretical coverage of the learning design.' Based on expert feedback, targeted revisions were carried out prior to empirical testing: sentence structures in four items were simplified, contextual cues were added to three Future Innovation items, and answer option formatting was standardized across all items. All 30 items were retained after revision. This iterative refinement process is consistent with best practices in instrument development (Clark & Watson, 2019), ensuring that content validity concerns were addressed systematically before proceeding to empirical validation. These revisions strengthened the alignment between the instrument items, the theoretical framework, and the characteristics of the target respondents. Consequently, the revised instrument was considered sufficiently robust to undergo empirical testing for further evaluation of its psychometric properties.

Content validity is widely recognized as a fundamental stage in instrument development because it establishes whether the items adequately represent the construct intended to be measured. This process relies on systematic expert judgment to evaluate the relevance, representativeness, and appropriateness of each item in relation to the assessment objectives (Almanasreh et al., 2019). In addition, contemporary instrument development frameworks emphasize that items should be derived directly from clearly defined constructs, organized logically, and adapted to the linguistic

characteristics, contextual background, and cognitive level of the target population before empirical validation is conducted (De Sá Leite et al., 2018; Henderson et al., 2025).

Expert review is not intended solely to determine the acceptability of an instrument but also to guide iterative refinement of the items. Revisions are commonly undertaken when expert agreement does not reach the predetermined criteria, when item wording is ambiguous, or when reviewers identify problems related to clarity and respondent interpretation (Anggara et al., 2023; Koskey et al., 2018). Such an iterative revision process enhances the alignment between theoretical constructs, assessment items, and respondents' understanding, thereby strengthening the quality of the instrument before proceeding to empirical psychometric evaluation (Henderson et al., 2025).

Empirical validity was assessed using Pearson's product-moment correlation between each item score and the total test score. With N = 30 and α = 0.05, the critical value r-table = 0.361. An item was declared valid if r-count ≥ r-table and p-value < 0.05. The complete results are presented in Table 4.

Table 4. Results of validity analysis

Item	r-count	r-table (N=30, α=0.05)	Sig. (p-value)	Decision
Item_1	0.494	0.361	0.006	Valid
Item_2	0.367	0.361	0.046	Valid
Item_3	0.516	0.361	0.004	Valid
Item_4	0.387	0.361	0.035	Valid
Item_5	0.713	0.361	0.000	Valid
Item_6	0.555	0.361	0.001	Valid
Item_7	0.440	0.361	0.015	Valid
Item_8	0.349	0.361	0.059	Not Valid
Item_9	0.426	0.361	0.010	Valid
Item_10	0.541	0.361	0.002	Valid
Item_11	0.335	0.361	0.070	Not Valid
Item_12	0.516	0.361	0.004	Valid
Item_13	0.462	0.361	0.010	Valid
Item_14	0.274	0.361	0.143	Not Valid
Item_15	0.484	0.361	0.007	Valid
Item_16	0.586	0.361	0.001	Valid
Item_17	0.355	0.361	0.055	Not Valid
Item_18	0.230	0.361	0.222	Not Valid
Item_19	0.701	0.361	0.000	Valid
Item_20	0.479	0.361	0.007	Valid
Item_21	0.558	0.361	0.001	Valid
Item_22	0.447	0.361	0.013	Valid
Item_23	0.426	0.361	0.019	Valid
Item_24	0.380	0.361	0.038	Valid
Item_25	0.643	0.361	0.000	Valid
Item_26	0.421	0.361	0.021	Valid
Item_27	0.407	0.361	0.026	Valid
Item_28	0.416	0.361	0.022	Valid
Item_29	0.392	0.361	0.032	Valid
Item_30	0.669	0.361	0.000	Valid

Empirical validity was assessed using Pearson product-moment correlation between each item score and the total test score. With N = 30 and α = 0.05, the critical value r-table = 0.361. An item was declared valid if r-count ≥ r-table and p-value < 0.05. Out of 30 items tested, 25 items were declared valid and 5 items were declared invalid (Items 8, 11, 14, 17, and 18). Valid items showed r-count values ranging from 0.367 (Item 2) to 0.713 (Item 5), indicating moderate to strong item-total correlations. The two highest correlations were found in Item 5 (r = 0.713, Problem Understanding indicator) and Item 19 (r = 0.701, Inference indicator), suggesting that these items are strong and consistent representatives of the FOCTS construct. The five invalid items all shared a common pattern: their r-count values fell below the r-table threshold and their p-values exceeded 0.05. Item 18 recorded the weakest correlation (r = 0.230; p = 0.222), followed by Item 14 (r = 0.274; p = 0.143), Item 11 (r = 0.335; p = 0.070), Item 17 (r = 0.355; p = 0.055), and Item 8 (r = 0.349; p = 0.059). These low correlations suggest that the five items failed to consistently measure the same

underlying construct as the rest of the instrument, likely due to ambiguous item wording, insufficient alignment between the item stem and the intended FOCTS indicator, or the influence of extraneous variables during the trial (Pillet et al., 2023). From a theoretical standpoint, validity refers to the degree to which an instrument measures what it is intended to measure (Ramadhan et al., 2024). The fact that 83.3% of items (25 out of 30) demonstrated significant item-total correlations indicates that the FOCTS construct is well-represented by the instrument as a whole. This is consistent with findings from comparable studies; (Aktoprak & Hursen, 2022) reported a validity rate of 82% in developing a critical thinking test for elementary students, affirming that a validity rate above 80% reflects strong instrument quality. The five invalid items were subsequently removed, and reliability analysis was conducted on the final 25-item version, in line with standard psychometric practice.

Internal consistency reliability was evaluated using Cronbach's Alpha on the 25 items that met the empirical validity criteria. This analysis was conducted to determine the consistency of the instrument in measuring the intended construct across all retained items. The result is presented in Table 5.

Table 5. Reliability statistics

Cronbach's Alpha	Number of items	Category
0.871	25	High Reliability

The analysis yielded $\alpha = 0.871$, which exceeds the commonly accepted minimum threshold of 0.70 for research instruments (Cerri et al., 2023) and surpasses the stricter threshold of 0.80 recommended for high-stakes diagnostic instruments. According to the reliability classification by Malapane & Ndlovu (2024), a Cronbach's Alpha value between 0.80 and 0.90 is categorized as high reliability, meaning the instrument consistently measures the FOCTS construct across similar respondent groups. This high reliability coefficient indicates that the 25 items exhibit strong inter-item homogeneityeach item contributes meaningfully and consistently to the measurement of the overall construct without introducing systematic measurement error.The attainment of $\alpha = 0.871$ is particularly noteworthy given the multidimensional nature of the FOCTS framework, which spans five distinct indicators (Problem Understanding, Analysis, Evaluation, Inference, and Future Innovation) at different Bloom's Taxonomy levels (C4 - C6). Achieving high internal consistency across these diverse cognitive demands suggests that the items, despite measuring different facets of future-oriented critical thinking, converge on a unified latent trait. This is theoretically coherent with the proposition that critical thinking, while multifaceted, constitutes a holistic and interconnected cognitive competency (Facione, 2011). Comparatively, similar studies in elementary science assessment have reported Cronbach's Alpha values ranging from 0.75 to 0.89. For instance, Series (2020) reported $\alpha = 0.82$ for a higher-order thinking skills (HOTS) test in elementary science, while Zulkipli (2020) obtained $\alpha = 0.79$ for a science reasoning instrument at the same level. The present instrument's reliability coefficient of 0.871 positions it favorably within this range, confirming its suitability as a reliable measurement tool. These findings suggest that the FOCTS diagnostic test can be expected to yield stable and replicable results when administered to students with comparable academic profiles, thereby supporting its use as a dependable assessment instrument in research and educational practice.

The uniqueness of this study lies in the development of a diagnostic instrument that specifically measures Future Innovation, namely the future-oriented and generative dimension of critical thinking at the C6 level according to Bloom's taxonomy. This aspect has received insufficient attention in research on critical thinking assessment at the elementary school level. Previous instruments have primarily focused on students' analytical and evaluative skills (Guntur et al., 2023; Mahoney et al., 2023), whereas this instrument aims to assess students' ability to generate innovative and future-relevant solutions through classroom-based objective tests. Furthermore, this study makes a methodological contribution by reporting content validity evidence separately for logical validity and face validity, thereby offering a more transparent evaluation of expert judgment during the instrument development process.

Several limitations of this study should be acknowledged. First, the sample size was relatively small N = 30 and drawn from a single school, which may limit the generalizability of the findings. Future research should involve a larger and more diverse sample to provide stronger evidence regarding the application of this instrument in various educational contexts. Second,

evidence of construct validity beyond item-total correlations was not tested in this study. Therefore, further research is needed to investigate the underlying structure of the instrument and to confirm the relationships among the five FOCTS indicators.

In the short term, the FOCTS diagnostic instrument provides teachers with a practical tool for identifying students' strengths and weaknesses based on five future-oriented indicators of critical thinking. The use of this instrument also highlights the importance of assessment literacy as part of teacher professionalism. Teachers need to be able to interpret diagnostic results and use them to improve teaching practices. Therefore, the future implementation of this instrument must be accompanied by professional development efforts that strengthen teachers' assessment competencies. Such professional development should emphasize not only technical skills in administering and scoring the instrument but also the pedagogical capacity to translate assessment evidence into targeted instructional interventions. Furthermore, sustained support through collaborative reflection and continuous monitoring can enhance teachers' confidence and consistency in using the instrument to foster students' future-oriented critical thinking.

4. Conclusion

This study has developed a diagnostic test instrument for FOCTS and examined its content validity, empirical validity, and reliability for use in fifth-grade elementary science learning. Overall, the instrument demonstrated adequate expert-judged content validity, a satisfactory proportion of empirically valid items, and high internal consistency reliability, indicating that the instrument shows promising psychometric properties as a foundation for diagnosing students' future-oriented critical thinking profiles. The principal novelty of this study lies in its explicit operationalization of the future-oriented, generative dimension of critical thinking into a measurable diagnostic instrument grounded in both logical and face validity evidence, addressing a gap not yet filled by existing critical thinking assessments at the elementary level. These results should nonetheless be interpreted as preliminary rather than conclusive evidence of the instrument's readiness for wide-scale use. The validation process was constrained by a small sample (N = 30) drawn from a single school, a mismatch between the intended fifth-grade target population and the sixth-grade trial sample, and the absence of construct validity evidence beyond item-total correlation. The instrument is therefore best characterized as a promising but preliminary diagnostic tool that requires further validation before broader implementation.

Acknowledgment

The authors would like to express their sincere gratitude to the principal, teachers, and students of the participating elementary school in Sukabumi for their cooperation in this research. The authors also thank the expert validators for their valuable feedback and acknowledge the support of Sukabumi Muhammadiyah University and Ahmad Dahlan University in collaborative study.

Declarations

Autor contribution : EK conceptualized the study, designed the FOCTS instrument, conducted the empirical testing, analyzed the data, and wrote the original draft of the manuscript. SS contributed to the theoretical framework development, reviewed and validated the instrument items, and provided critical feedback on the manuscript. I. supervised the overall research process, verified the statistical analyses, and revised the final manuscript. All authors reviewed and approved the final version of the manuscript.

Funding statement : This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest : The authors declare no conflict of interest.

Additional information : No additional information is available for this paper.

References

Aini, N., & Putri, R. T. H. (2024). Validity of E-LKPD to train students' critical thinking abilities. *APPLICATION: Applied Science in Learning Research, 4*(1), 33-39. https://doi.org/10.32764/application.v4i1.4699

Aktoprak, A., & Hursen, C. (2022). A bibliometric and content analysis of critical thinking in primary education. *Thinking Skills and Creativity*, *44*, 101029. https://doi.org/10.1016/j.tsc.2022.101029

Almanasreh, E., Moles, R., & Chen, T. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy, 15*, 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Alshaikh, A. A. N. (2024). Evaluating future-oriented thinking skills in educational communication of students at Prince Sattam bin Abdulaziz University. *Profesional de La Información*, *33*(4), e330415. https://doi.org/10.3145/epi.2024.ene.0415

Anggara, D. S., & Abdillah, C. (2023). Content validity *analysis of literacy assessment instruments. Jurnal Cakrawala Pendidikan, 42*(2), 447-459.. https://doi.org/10.21831/cp.v42i2.55900

Aristanti, A. P., & Fatayan, A. (2024). The effect of the UbD-Based problem based learning model on the critical thinking skills of grade IV students in IPAS subject at elementary school. *Mimbar Sekolah Dasar*, *11*(2), 268–280. https://doi.org/10.53400/mimbar-sd.v11i2.71794

Badenes-Ribera, L., Silver, N. C., & Pedroli, E. (2020). Editorial: Scale development and score validation. *Frontiers in Psychology, 11,* 799. https://doi.org/10.3389/fpsyg.2020.00799

Care, E., & Griffin, P. (2022). An approach to assessment of collaborative problem solving. *Res. Pract. Technol. Enhanc. Learn., 9*, 367-388. https://doi.org/10.58459/rptel.2014.9367-388

Cerri, L. Q., Justo, M. C., Clemente, V., Gomes, A. A., Pereira, A. S., & Marques, D. R. (2023). Insomnia Severity Index: A reliability generalisation meta-analysis. *Journal of Sleep Research*, *32*(4), e13835. https://doi.org/10.1111/jsr.13835

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

De Sá Leite, S., Áfio, A. C. E., Carvalho, L. V., Da Silva, J. M., Almeida, P. C., & Pagliuca, L. (2018). Construction and validation of an educational content validation instrument in health. *Revista Brasileira de Enfermagem, 71,* 1635-1641. https://doi.org/10.1590/0034-7167-2017-0648

Dilekçi, A., & Karatay, H. (2023). The effects of the 21st century skills curriculum on the development of students' creative thinking skills. *Thinking skills and creativity*, *47*, 101229. https://doi.org/10.1016/j.tsc.2022.101229

Ennis, R. H. (2018). Critical thinking across the curriculum: A vision. *Topoi*, *37*(1), 165–184. https://doi.org/10.1007/s11245-016-9401-4

González-Pérez, L. I., & Ramírez-Montoya, M. S. (2022). Components of education 4.0 in 21st century skills frameworks: Systematic review. *Sustainability*, *14*(3), 1493. https://doi.org/10.3390/su14031493

Guntur, M., Sholeha, H. H., Kurniawati, I., & Astuti. (2023). Analysis of students' critical thinking patterns on critical thinking ability: The case of elementary school. *Widya Accarya*. *14*(2), 199-210. https://doi.org/10.46650/wa.14.2.1469.199-210

Henderson, T., Godwin, A., & Carberry, A. (2025). Supporting fair use in context of instruments in engineering education research: A renewed conversation about validity. *Journal of Engineering Education, 115*(1), e70041. https://doi.org/10.1002/jee.70041

Koskey, K. L. K., Sondergeld, T., Stewart, V., & Pugh, K. J. (2018). Applying the mixed methods instrument development and construct validation process: The transformative experience questionnaire. *Journal of Mixed Methods Research, 12*, 122-95. https://doi.org/10.1177/1558689816633310

Kurniasari, E., Sugiyarto, S., Ishafit, I., & Thobirin, A. (2026). A data-driven comparison of game based and problem based learning in primary education. *International Journal of Innovative Research and Scientific Studies*, *9*(3), 106–117. https://doi.org/10.53894/ijirss.v9i3.11397

Liu, J., Liu, Z., Wang, C., Xu, Y., Chen, J., & Cheng, Y. (2024). K-12 students' higher-order thinking skills: Conceptualization, components, and evaluation indicators. *Thinking Skills and Creativity*, *52*, 101551. https://doi.org/10.1016/j.tsc.2024.101551

Loyens, S. M. M., van Meerten, J. E., Schaap, L., & Wijnia, L. (2023). Situating higher-order, critical, and critical-analytic thinking in problem- and project-based learning environments: A systematic review. *Educational Psychology Review*, *35*(39), 1-44. https://doi.org/10.1007/s10648-023-09757-x

Mahoney, B. B., Oostdam, R. R., Nieuwelink, H. H., & Schuitema, J. J. (2023). Learning to think critically through Socratic dialogue: Evaluating a series of lessons designed for secondary vocational education. *Thinking Skills and Creativity*, *50*, 101422. https://doi.org/10.1016/j.tsc.2023.101422

Malapane, T. A., & Ndlovu, N. K. (2024). Assessing the reliability of likert scale statements in an e-commerce quantitative study: A cronbach alpha analysis using SPSS statistics. *2024 Systems and Information Engineering Design Symposium (SIEDS)*, 90–95. https://doi.org/10.1109/SIEDS61124.2024.10534753

Morad, S., Ragonis, N., & Barak, M. (2021). The validity and reliability of a tool for measuring educational innovative thinking competencies. *Teaching and Teacher Education*, *97*, 103193. https://doi.org/10.1016/j.tate.2020.103193

Musfidah, H., Bachri, B., Susarno, L. H., & Rofiah, K. (2023). Evaluation of merdeka belajar curriculum from the views of high school students: Implementation of differentiated learning in economics. *International Journal of Emerging Research and Review*, *1*(4), 000052. https://doi.org/10.56707/ijoerar.v1i4.52

Pillet, J., Carillo, K. D., Vitari, C., & Pigni, F. (2023). Improving scale adaptation practices in information systems research: Development and validation of a cognitive validity assessment method. *Information Systems Journal*, *33*(4), 842–889. https://doi.org/10.1111/isj.12428

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

Sarwanto, S., Fajari, L. E. W., & Chumdari, C. (2020). Open-ended questions to assess critical-thinking skills in indonesian elementary school. *International Journal of Instruction, 14*(1), 615-630. https://doi.org/10.29333/iji.2021.14137a

Schäfer, J., Reuter, T., Leuchter, M., & Karbach, J. (2024). Executive functions and problem-solving—The contribution of inhibition, working memory, and cognitive flexibility to science problem-solving performance in elementary school students. *Journal of Experimental Child Psychology*, *244*, 105962. https://doi.org/10.1016/j.jecp.2024.105962

Series, C. (2020). *Higher order thinking skills (HOTS) -oriented e-module in electric circuit*, *1521*, 1-7 https://doi.org/10.1088/1742-6596/1521/2/022027

Suhendar, S., Widodo*, A., Solihat, R., Riandi, R., & Chaw, E. P. (2025). Teaching strategies for enhancing future thinking skills in secondary education: A systematic literature review. *Jurnal IPA & Pembelajaran IPA*, *9*(1), 20–33. https://doi.org/10.24815/jipi.v9i1.42446

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., Mourey, F., Feybesse, C., Sundquist, D., & Lubart, T. (2023). Creativity, Critical Thinking, Communication, and Collaboration: Assessment, Certification, and Promotion of 21st Century Skills for the Future of Work and Education. *Journal of Intelligence*, *11*(3), 54. https://doi.org/10.3390/jintelligence11030054

Zulkipli, Z. A. (2020). *Identifying scientific reasoning skills of science education students*. *16*(3), 275-280. https://doi.org/10.24191/ajue.v16i3.10311