Proceedings of the International Conference on Education, Teaching, and Learning, Vol. 4, 2025, pp. 1-17
Copyright © 2025 iConferences
ISSN: 2820-2155
DOI: https://doi.org/10.32789/edcon.2025.4101

Development of an Online Physics Assessment Instrument for Computational Thinking Skills Integrated with the Batik-Making Process of *Parang* Motifs (PhyCoBa)

[1]Ayu Mar'ati Barokatun N, [1]Nisya Indriyanti Irawan, [1]Supahar, [1]Bonita Padang, [1]Rida Siti Nuraini Mahmudah

[1]*Yogyakarta State University, Indonesia,* ayumarati.2024@student.uny.ac.id; nisyaindriyanti.2024@student.uny.ac.id; supahar@uny.ac.id; bonitapadang.2024@student.uny.ac.id; rida@uny.ac.id

DEVELOPMENT OF AN ONLINE PHYSICS ASSESSMENT INSTRUMENT FOR COMPUTATIONAL THINKING SKILLS INTEGRATED WITH THE BATIK-MAKING PROCESS OF *PARANG* MOTIFS (PhyCoBa)

Abstract: Assessment instruments must accurately measure complex 21st-century competencies while maintaining contextual relevance. This study developed PhyCoBa (Physics, Computational Thinking, and Batik), an innovative online assessment instrument designed to measure physics Computational Thinking (CT) skills by integrating the ethnoscience context of the Parang batik-making process. Using a modified development model from Wilson, Oriondo, and Antonio, a 20-item multiple-choice test was constructed and validated by physics education experts and practitioners. The instrument was trialed with 165 Grade XI students in the Yogyakarta Special Region, selected via stratified purposive sampling across high, medium, and low achievement levels. Data analysis using the Rasch 1-Parameter Logistic (1-PL) model showed that all items achieved a proper fit. The instrument demonstrated high technical quality with a reliability of 0.95 and an item difficulty index ranging from -1.36 to 1.72, confirming that all items fall within a good category. PhyCoBa offers a valid and culturally engaging solution for educators to monitor and support the development of computational thinking within physics curricula. By bridging traditional cultural heritage with modern digital assessment, this instrument facilitates more authentic and meaningful learning evaluations in the science classroom.

Keywords: Physics, assessment instruments, computational thinking, batik-making process

Introduction

Education in the 21st century necessitates cultivating skills relevant to global challenges and the Fourth Industrial Revolution. One of the most pivotal competencies in this landscape is Computational Thinking (CT) (Kaya *et al*., 2025; Mariana & Kristanto, 2023). Recognized by international frameworks such as the ISTE standards and the PISA 2022 framework, CT is not confined to the domain of computer science but encompasses a series of thought processes for systematically solving complex problems (Tolba, 2025). CT comprises four primary pillars: decomposition (breaking down a complex problem into smaller, manageable parts), pattern recognition (identifying similarities or trends), abstraction (focusing on essential information while disregarding irrelevant details), and algorithmic design (developing step-by-step solutions) (Mariana & Kristanto, 2023; Tolba, 2025; ZSOLDOS-MARCHIȘ & BÁLINT-SVELLA, 2023). Research by Herlina, Abdurrahman & Umam (2025) found that without deliberate pedagogical design, students often fail to recognize or apply CT in the science domain. Therefore, integrating CT across disciplines, including physics, is crucial for preparing students to think logically, structurally, and innovatively when confronting real-world problems.

Physics education is often perceived by students as abstract and challenging, requiring deep conceptual understanding and mathematical proficiency. However, physics is essentially the science of modeling natural phenomena, a process that inherently aligns with the pillars of

CT (Yin *et al*., 2020). Physics students must deconstruct complex phenomena (decomposition), identify underlying laws or principles (pattern recognition), create idealized models by omitting minor factors (abstraction), and formulate problem-solving steps (algorithms) (Hutchins *et al.*, 2020; Orban & Teeling-Smith, 2020). Despite this intrinsic alignment, the implementation of physics instruction that explicitly cultivates CT remains limited, as many assessment methods continue to emphasize rote memorization of formulas and procedural problem-solving, failing to foster underlying thought processes**.**

In line with the paradigm shift in learning, assessment paradigms must also evolve. To effectively measure students' mastery of CT skills within the context of physics, specially designed assessment instruments are required (Zakwandi & Istiyono, 2023). Consequently, the development of digital-based assessment instruments presents a relevant solution (Lu *et al.*, 2022). Online assessment offers significant advantages, including unparalleled flexibility and broader accessibility for students. It also provides the capacity for instantaneous feedback, which is crucial for the iterative nature of CT skill development (Zapata-Caceres *et al.*, 2021). Most importantly, digital platforms can present interactive and authentic problem-solving scenarios that more effectively represent the complexities and challenges inherent in CT-based tasks than traditional methods (Eloy *et al.*, 2022; Lai, 2022). These dynamic assessments enable a more in-depth and comprehensive evaluation of students' thought processes and problem-solving strategies.

To enhance student engagement and instructional relevance, an approach that integrates local wisdom into science, known as ethnoscience, has proven effective. Batik, a UNESCO Intangible Cultural Heritage from Indonesia, offers a rich context for exploring concepts in both physics and computational thinking (Marina *et* al., 2025; Mukarom *et al*., 2024). The process of creating batik, particularly the Parang motif, is laden with relevant principles. The geometric and repetitive Parang motif embodies physics concepts such as transformations, symmetry, and wave patterns *(Susanto et al*., 2025). Furthermore, its creation process—from fabric preparation and pattern drawing to wax application (canting) and dyeing—constitutes a systematic and algorithmic procedure.

A significant research gap currently exists: the lack of an integrated, valid, and reliable assessment instrument designed to measure CT skills within a physics context that is also culturally engaging for Indonesian students (Brinkman *et al.*, 2017). Existing instruments tend to focus on a single facet, such as general physics assessment, CT within a programming context, or culture-based learning that lacks structured CT evaluation (Purnomo *et al*., 2022). This study addresses these gaps by developing the PhyCoBa instrument—an acronym for Physics, Computational thinking, and Batik—integrated with the Parang motif-making process.

Therefore, this research addresses two primary questions: (1) What is the development process of the assessment instrument? and (2) what are its psychometric characteristics? This

study proceeds with the hypothesis that a culturally contextualized digital assessment instrument, grounded in the Rasch 1-PL model, can accurately and reliably differentiate students' computational thinking abilities while simultaneously fostering an appreciation for national cultural heritage.

Methods

Research Design and Procedure

This quantitative development research adapted a modified framework based on the Wilson and Oriondo & Antonio models. The development process encompassed three specific phases: planning, construction, and field trial. The procedural flow for developing the test instrument is illustrated in Figure 1.

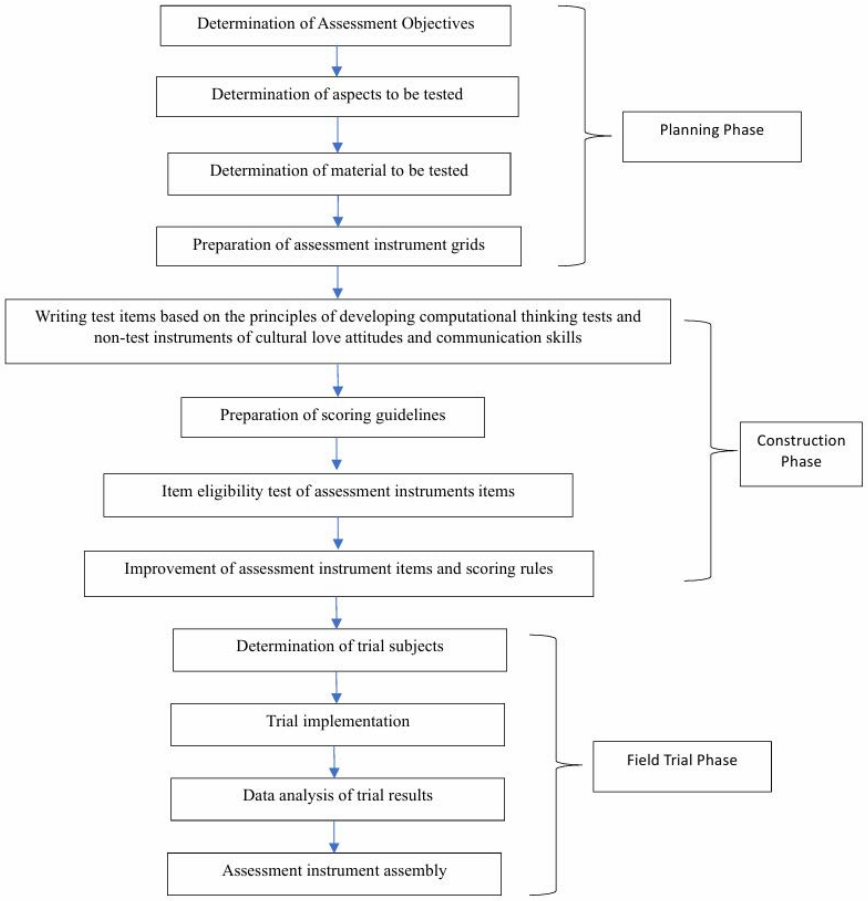


Figure 1 Flow of Assessment Instrument Development

The procedure commenced with the planning phase, which involved determining assessment objectives, specifying measured aspects **of** computational thinking, and preparing a comprehensive blueprint to ensure content validity. This was followed by the construction phase, where the blueprint was operationalized into test and non-test items and scoring

guidelines were simultaneously developed**.** To ensure high-quality content, the initial draft underwent an expert judgment process. The instrument was validated by a panel of physics education experts and practitioners who provided feedback on clarity and relevance. Feedback from these validators was used to calculate Aiken's V statistics and perform preliminary improvements before the field trial**.**

Research Subject

The study's trial sample comprised 165 Grade XI (Phase F) students from three senior high schools in the Special Region of Yogyakarta. Grade XI was selected because these students had already covered the prerequisite material, "Quantities and Measurements," in Grade X, ensuring that their responses reflected their computational thinking abilities rather than a lack of subject knowledge**.** The sampling technique employed was stratified purposive sampling. Schools were stratified into high, medium, and low achievement categories based on their 2022 average UTBK results to ensure sample heterogeneity. Ethical considerations were strictly maintained during the trial; formal permission was obtained from the participating schools, and informed consent was secured from the students before trial implementation**.**

Data Analysis

The assessment instrument utilized a multiple-choice format, with items scored dichotomously: 1 for a correct answer and 0 for an incorrect answer. The Rasch 1-Parameter Logistic (1-PL) model was selected for calibration because it is specifically suited for such response data. The 1-PL model was justified by its ability to transform ordinal raw scores into interval-scale measurements (logits), providing an objective measurement. Furthermore, a sample size of 165 was considered adequate to produce stable and accurate item parameter estimates, which is a critical requirement for this model**.**

Data analysis evaluated several psychometric properties based on established criteria: First, content validity was assessed using Aiken's V formula (Aiken, 1985); an item was considered "Highly feasible" if the V-value fell within the range of $0.8 < V \leq 1.0$. Second, construct validity was evaluated using item fit to the Rasch model. Based on the criteria of Adams & Khoo, (1996) An item or person was considered to fit (valid) if the INFIT MNSQ value was within the range of 0.77 to 1.30. Furthermore, instrument reliability was determined using the reliability coefficient (R). Referring to the criteria by Gliem & Gliem, (2003) The instrument was categorized as "Highly Reliable" if it had an R-value of $R \geq 0.80$. The item difficulty level (b) was analysed using the criteria from Hambleton & Swaminathan (1985), where items were classified as “Easy” ($-2 \leq b < -1$), “Moderate” $-1 \leq b \leq 1$), and “Difficult ($1 < b \leq 2$). Finally, the Test Information Function (TIF) and Standard Error of Measurement (SEM) were examined to confirm the instrument's accuracy across various student ability levels.

Table 1 Aiken's V Values

Score Range	Category
$0{,}8 < V \leq 1$	Highly Feasible
$0{,}6 < V \leq 0{,}8$	Feasible
$0{,}4 < V \leq 0{,}6$	Not Feasible
$0{,}2 < V \leq 0{,}4$	Highly Not Feasible

Table 2 INFIT MNSQ Values

INFIT MNSQ	Parameters
$> 1{,}30$	Invalid
$0{,}77 - 1{,}30$	Valid
$< 0{,}77$	Invalid

Table 3 Reliability Coefficient

Reliability Coefficient	Category
$R \geq 0{,}80$	Highly Reliable
$0{,}60 < R \leq 0{,}80$	Reliable
$0{,}40 < R \leq 0{,}60$	Fairly Reliable
$0{,}20 < R \leq 0{,}40$	Somewhat Reliable
$R \leq 0{,}20$	Less Reliable

Table 4 Threshold Values

Thresholds Values	Difficulty Levels
$b > 2$	Very Difficult
$1 < b \leq 2$	Difficult
$-1 \leq b \leq 1$	Moderate
$-2 \leq b \leq -1$	Easy
$b < -2$	Very Easy

Results and Discussion

Results

The development of the Physics, Computational Thinking, and Batik (PhyCoBa) assessment instrument resulted in 20 multiple-choice items covering dimensions of physics (quantities and measurements) and four CT pillars: decomposition, pattern recognition, abstraction, and algorithmic thinking. To ensure the rigor of the psychometric analysis, the following criteria were established.

Table 5 Distribution of Computational Thinking Skills Test Items

Aspect	Indicators	No Item	V-Aiken	Category
Decomposition	Using knowledge in real situations	1	1.00	Highly Recommended
	Application of understanding to	2	1.00	Highly

Aspect	Indicators	No Item	V-Aiken	Category
	identify units			Recommended
	Analysing each statement to evaluate its correctness	3	0.88	Highly Recommended
	Connecting a dimensional formula with a physical quantity	4	1.00	Highly Recommended
	Determining the dimension based on the known quantity	5	0.88	Highly Recommended
Pattern Recognition	Grouping quantities that have the same dimension	6	0.88	Highly Recommended
	Identifying the pattern of the width of batik wax lines	7	1.00	Highly Recommended
	Concluding the pattern recognizable from the dimensions of quantities	8	1.00	Highly Recommended
	Concluding the most accurate measurement	9	0.88	Highly Recommended
	Concluding the measuring instrument for density	10	0.88	Highly Recommended
Abstraction	Examining/analysing important information	11	0.88	Highly Recommended
	Breaking down measurement results according to the rules of significant figures	12	1.00	Highly Recommended
	Combining data and adding them up	13	1.00	Highly Recommended
	Determining the number of significant figures	14	1.00	Highly Recommended
	Determining the simplification of number notation	15	1.00	Highly Recommended
Algorithmic Thinking	Formulating systematic steps in measurement	16	1.00	Highly Recommended
	Formulating systematic steps for reading a measuring instrument	17	1.00	Highly Recommended
	Determining the reported result of a measuring instrument and its uncertainty	18	1.00	Highly Recommended
	Determining the reading of a measuring instrument	19	1.00	Highly Recommended
		20	1.00	Highly Recommended

The distribution of the computational thinking (CT) skills test instrument, detailed in Table 5, is meticulously structured to assess four core components of CT: Decomposition, Pattern Recognition, Abstraction, and Algorithmic Thinking. The instrument comprises a total of 20 items, with five items specifically dedicated to evaluating each of the four designated aspects,

ensuring a balanced representation across the construct. The indicators tested are contextually relevant to physics and measurement, such as utilizing knowledge in real situations and determining dimensions (Decomposition); grouping quantities with similar dimensions and concluding patterns (Pattern Recognition); applying rules of significant figures and number simplification (Abstraction); and formulating systematic steps for measurement and reporting uncertainty (Algorithmic Thinking). This balanced design ensures that the test provides a comprehensive assessment of students' abilities to apply CT principles within a scientific domain.

Furthermore, the content validity of the instrument, evaluated using Aiken's V statistic, confirms the high quality and suitability of every item. The calculated V-Aiken values for all 20 test items exhibit a narrow, high-range spectrum, specifically spanning from 0.88 to 1.00. According to the established content validity criteria by Aiken, (1985), any instrument item that falls within the score range of $0.8 < V \leq 1$ is classified as "Highly feasible". Consequently, the universal range of V-Aiken values obtained places all 20 items in the Highly feasible category. The unanimous classification of all items as "Highly Recommended" underscores the robust consensus among the expert panel regarding the content relevance and clarity of the test, firmly validating the instrument for measuring the four dimensions of computational thinking skills.

Unidimensional Assumption Test

Before proceeding with advanced psychometric analyses, such as Item Response Theory (IRT) or Rasch modeling, a crucial preliminary step is validating the unidimensionality assumption. Unidimensionality is a fundamental requirement in measurement theory, asserting that all items within a test or scale must predominantly measure a single underlying psychological dimension, trait, or construct (Version & Dongen, 2020). In essence, it ensures that the variation in test scores is primarily attributable to a difference in a single ability, rather than a mix of several uncorrelated attributes. Although it is rare for an instrument to be perfectly unidimensional due to the complexity of human cognition, the assumption is considered met if one dominant factor is found to explain a significantly larger portion of the total variance than any other factor (Ayanwale *et al*., 2020). This assumption is typically evaluated by using statistical techniques like Principal Component Analysis (PCA) of the residuals or Exploratory Factor Analysis (EFA) (Ayanwale *et al*., 2020), where the ratio and magnitude of the first eigenvalue relative to the subsequent eigenvalues serve as the primary evidence for the existence of a single, powerful latent variable.

Table 6 Initial Eigenvalues of Components

Item	Eigenvalues		
	Total	% of Variance	Cumulative %
1	3.698	18.492	18.492
2	1.566	7.829	26.321
3	1.362	6.809	33.130
4	1.316	6.582	39.172
5	1.110	5.549	45.261
6	1.066	5.330	50.591
7	1.028	5.139	55.731

The unidimensionality assumption is a cornerstone of rigorous psychometric evaluation, asserting that the items within a measurement instrument are predominantly attributable to a single underlying latent construction. This premise is vital for the valid interpretation of test scores. The unidimensionality of the instrument was assessed using EFA on the item responses, as detailed in the Total Variance Explained table derived from the SPSS output.

The analysis initially identified multiple factors with eigenvalues exceeding the Kaiser criterion (eigenvalues > 1). Specifically, seven components were observed to have initial eigenvalues greater than unity, explaining a cumulative variance of 55.731%. However, for the assumption of practical unidimensionality to be met, the first factor must demonstrate pronounced dominance over all subsequent factors. The results clearly indicate that the first factor possesses a remarkably high eigenvalue of 3.698, which accounts for 18.492% of the total variance

Crucially, the eigenvalue of the first factor (3.698) is more than double the eigenvalue of the second factor (1.566), and significantly higher than the eigenvalue of the third factor (1.362). This substantial disparity and the isolation of the first factor as the primary source of variance provide strong empirical evidence for the existence of a single, highly dominant latent variable driving the item responses. Although the presence of multiple factors with eigenvalues greater than 1 suggests that some residual variance is captured by minor dimensions, the overwhelming dominance of the first factor leads to the conclusion that the test is sufficiently unidimensional for its intended purpose (Ocy & Sarifah, 2025). Therefore, the requirement of unidimensionality is fulfilled, substantiating the use of the instrument to measure the overall construction of Computational Thinking skills.

Local Independence Assumption Test

Beyond the unidimensionality assumption, the local independence assumption is a fundamental requirement in measurement models based on Item Response Theory (IRT) and Rasch. This assumption states that, after the underlying latent factor (the measured construct, e.g., computational thinking skills) has been controlled for or partialled out, a respondent's

answer to one item should not be correlated with their answer to any other item. Practically, local independence ensures that the remaining variance (residual) in the data does not contain any significant additional dimensions (Ayanwale *et al*., 2020). Violation of this assumption indicates that factors other than the targeted construct are influencing respondent answers on certain item pairs. Theoretically, the local independence assumption is automatically satisfied if the unidimensionality assumption has been met (Ayanwale *et al.*, 2020). Since the assessment instrument developed in this study has demonstrated fulfillment of the unidimensionality assumption (as described in the previous section through the dominance of the first factor), it is deductively concluded that the assessment instrument for measuring computational thinking skills has also met the local independence assumption.

Instrument Item Goodness Fit and Level of Difficulty

Testing the determination of the fit of each item on the model follows the rules of (Adams & Khoo, 1996), namely, an item fits the model if the INFIT MNSQ value is between 0.77 and 1.30. With the limits of acceptance of items using INFIT MNSQ or fit according to the model (between 0.77 and 1.30) and using threshold or item difficulty level with a limit of -2.0 to 2.0, the items that are suitable to fulfill the goodness of fit are obtained.

Table 7 MISFIT ORDER

Item	INFIT MNSQ	Category (Referring to Table 2)	Threshold	Category (Referring to Table 4)
1	0.96	Valid	0.77	Moderate
2	1.05	Valid	-1.26	Easy
3	0.86	Valid	-0.61	Moderate
4	0.87	Valid	0.23	Moderate
5	0.92	Valid	-0.17	Moderate
6	1.07	Valid	1.20	Difficult
7	0.96	Valid	0.68	Moderate
8	0.81	Valid	-0.17	Moderate
9	0.91	Valid	1.72	Difficult
10	1.02	Valid	-0.05	Moderate
11	1.15	Valid	1.14	Difficult
12	0.99	Valid	0.59	Moderate
13	0.91	Valid	-1.00	Moderate
14	1.02	Valid	-0.91	Moderate
15	1.17	Valid	-0.11	Moderate
16	0.92	Valid	-1.36	Easy
17	1.27	Valid	-0.61	Moderate
18	1.01	Valid	-1.04	Easy
19	0.97	Valid	0.65	Moderate
20	1.07	Valid	0.32	Moderate

Based on the data in Table 7, all 20 items in the PhyCoBa instrument demonstrate an INFIT MNSQ value within the acceptable range of 0.77 to 1.30. Item 8 showed the lowest INFIT MNSQ value (0.81), while Item 17 showed the highest (1.27). Since no items fell outside these boundaries, it can be concluded that all items are psychometrically valid and contribute consistently to the measurement of the underlying construct: computational thinking skills integrated with the batik-making process.

In terms of specific difficulty levels, items such as R16 (-1.36), R2 (-1.26), and R18 (-1.04) were categorized as easy, which makes them highly suitable for accurately measuring students with lower ability levels. Meanwhile, the majority of the test items, including R1, R3, and R5, fell within the moderate range ($-1 \leq b \leq 1$), aligning well with the average computational thinking ability of Grade XI students. Furthermore, items such as R9 (1.72), R6 (1.20), and R11 (1.14) exhibited great difficulty levels, effectively providing discrimination for students with superior or above-average computational thinking skills. Since all 20 items remained within the ideal difficulty index limits of -2.0 to 2.0, it can be concluded that the PhyCoBa instrument possesses robust psychometric quality and is fully qualified to serve as a reliable tool for assessing students' computational thinking skills in a physics context.

Reliability

The results of the reliability analysis of the computational thinking assessment instrument can be seen in Table 8 below:

Table 8 Reliability Test of Computational Thinking Assessment Instrument

Reliability	Reliability Value's	Category (Referring to Table 3)
Person Reliability	0.72	Reliable
Item Reliability	0.95	Highly Reliable

The reliability of computational thinking questions based on the results of the analysis is reliable. Based on rasch model, the reliability value can be shown based on Pearson's Reliability and Item Reliability. Person reliability is 0.72, which shows the consistency of individual responses. The value of 0.74 is acceptable but can be increased for better consistency. The item reliability is 0.95, indicating excellent reliability, meaning that the items are consistent in measuring the underlying characteristics or abilities across respondents. This output indicates a reliable and valid model with good item separation; however, there is room for improvement in personnel reliability and individual response fit (Outfit MNSQ > 2.0). The high reliability and item separation values indicate that the items are effective in distinguishing between different individual abilities.

Item Characteristic Curve (ICC)

Each curve in the graph represents one measured item. This ICC shows how the probability of answering correctly changes as the individual's ability level changes. The ideal curve is sharp and shows a clear increase in probability with increasing ability.

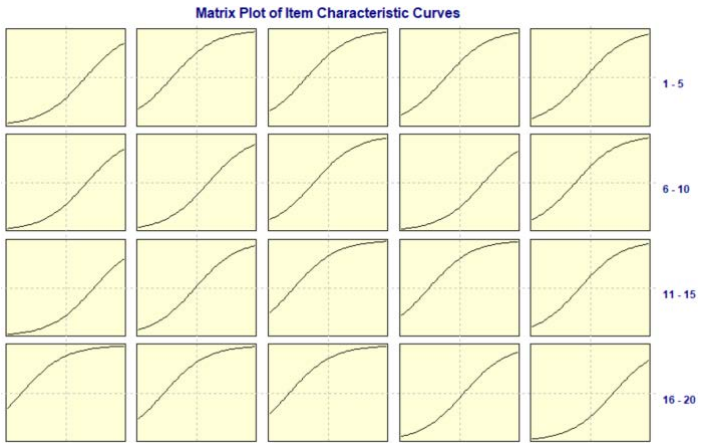


Figure 2 Item Characteristic Curve Computational Thinking Ability

The Item Characteristic Curve (ICC) graph shown in Figure 3 above illustrates the relationship between an individual's ability level or characteristics (usually measured on the X-axis) and the probability of success in answering the item (Y-axis) on each item measured. Items that have a sharper curve are better at distinguishing between low and high-ability individuals. This indicates that the item is quite effective in measuring the ability in question. Based on Figure 4, it can be concluded that the ICC for items number 1 to 20 tends to be sharp, so that the items are quite effective in distinguishing individuals with low and high abilities. ICC is also related to the level of item difficulty; the higher the level of item difficulty, the sharper the ICC will be, and vice versa. From Figure 4, the sharpest graph is owned by item 9, and the most sloping is owned by item 16. This is appropriate when referring to the data in Figure 2 because, based on Figure 2, item 16 has the lowest difficulty level (-1.36), and item 9 has the highest difficulty level (1.72).

Item Information Function and Standard Error Measurement (SEM)

The information curve shows the information provided by the test for different levels of scale scores. It indicates the extent to which the test provides useful information about an individual's ability at each level of ability tested.

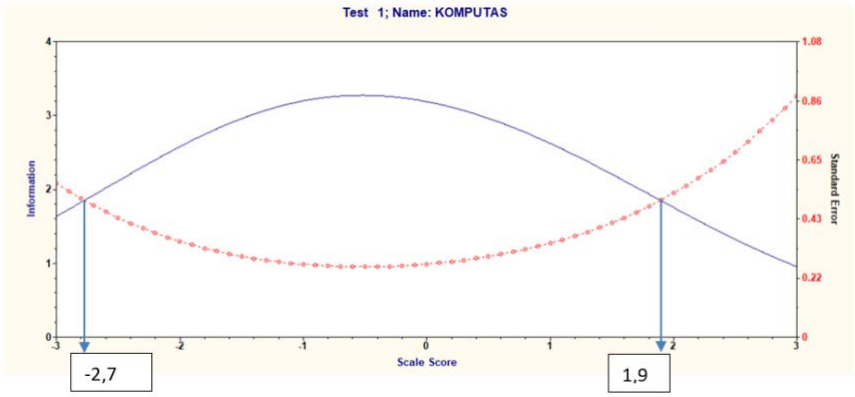


Figure 3 Information Function and SEM

The analysis results obtained information on the function and standard error measurement (SEM). Based on the information function and SEM stated in Figure 5, the test is suitable for students with computational thinking ability in the medium category, namely $-2,7 \leq \theta \leq 1,7$.

Discussion

For an assessment instrument to be effective, it must be both valid and reliable. Validity ensures the instrument accurately measures what it's intended to measure. This can be established through content validity, which is a rational analysis of the test's relevance to the topic. Reliability, on the other hand, refers to the instrument's ability to produce consistent results over time (Yaniawati, P., Maat, S. M., Supianti, I. I., & Fisher, 2022). Beyond being valid and reliable, a good assessment instrument should also have strong item characteristics. These characteristics, which include differentiability (how well an item distinguishes between high and low-performing students), difficulty, and apparent guessing (the likelihood of a student getting an answer right by chance), can be analyzed using either classical test theory or item response theory (IRT) (Crocker & Algina, 2008). The provided text notes that using IRT can reveal an instrument's characteristics and help identify which items are good and which ones need improvement.

The analysis of a 20-item assessment instrument for computational thinking skills indicates that it is a high-quality and reliable tool. All 20 items fit the model well, as evidenced by their INFIT MNSQ values ranging from 0.77 to 1.30 (Adams & Khoo, 1996). The items also have an ideal difficulty level, falling between -1.36 and 1.72, which means there are no questions that are too easy or too difficult for the respondents. The instrument's reliability is also strong, with an item reliability, indicating good internal consistency. Its ability to discriminate between individuals with low and high abilities is effective, as shown by high reliability and item separation values and sharp ICC curves. Based on the information function and SEM, this instrument is particularly suitable for assessing students with an average level of

computational thinking ability. Overall, all 20 items are considered good measures of computational thinking skills.

While the PhyCoBa instrument demonstrates excellent item reliability (0.95), a critical reflection reveals that Person Reliability remains at a moderate level (0.74). This suggests that while the items themselves are stable, the consistency of individual student responses could be further improved, potentially by increasing the number of items or refining the distractors to better capture varied student logic. Furthermore, the Information Function and SEM curve (Figure 3) indicate that this test is most effective for students within the medium-ability category ($-2.7 \leq \theta \leq 1.7$). Consequently, educators should be cautious when using this specific version of PhyCoBa for students at the extreme ends of the ability spectrum, as the measurement error increases significantly for very high or very low achievers.

The integration of the Parang batik-making process as an ethnoscience context represents a significant departure from many international CT assessment tools. Most global instruments, such as those discussed by Kaya *et al.* (2025), often focus on pure programming environments or decontextualized logic puzzles. In contrast, PhyCoBa aligns with the modern pedagogical shift toward situated CT, where computational processes are embedded in real-world cultural practices. Like the C2STEM framework developed by Saritepeci & Durak, (2022), which integrates physics with CT through modeling, PhyCoBa proves that cultural heritage can serve as a powerful cognitive bridge. The success of the Parang motif context in this study confirms that ethnoscience does not diminish psychometric rigor but rather provides a familiar "algorithmic" structure that helps students navigate abstract physics concepts more effectively.

Conclusions

Based on the research objectives and data analysis, it can be concluded that this study successfully addressed both research questions. First, the Online Physics Assessment Instrument for Computational Thinking (PhyCoBa) was developed by adapting the Wilson and Oriondo & Antonio models through three systematic phases: planning, field trial, and final assembly**.** Second, regarding its psychometric characteristics, PhyCoBa was proven to be valid, as all 20 items fit the Rasch 1-Parameter Logistic (1-PL) model with INFIT MNSQ values ranging between 0.77 and 1.30. The instrument also demonstrated high reliability, with a Cronbach's Alpha of 0.74 and an item reliability of 0.95. Furthermore, all items exhibited appropriate difficulty levels, ranging from -1.36 to 1.72, indicating that the instrument is most effective for measuring students with moderate-level computational thinking skills.

Practically, this instrument serves as a valuable tool for teachers to implement formative assessments, allowing them to monitor students' thinking processes more objectively. For curriculum developers, this study provides a model for integrating local cultural values—

specifically ethnoscience—into science materials to make learning more contextual and engaging. Future research should focus on further validating this instrument across different educational levels and wider geographic regions to enhance its generalizability. Additionally, there is a significant opportunity to develop assessment tools for other physics topics that integrate various forms of local wisdom to further enrich the science curriculum.

Acknowledgments

We would like to express our sincere gratitude to the Indonesian Endowment Fund for Education (LPDP) for their generous financial support through the LPDP Scholarship Program. This scholarship has been instrumental in funding this research/study and has provided us with the opportunity to pursue our academic goals. We are truly thankful for their belief in our work and their commitment to supporting education and research.

References

Adams, R. J., & Khoo, S.-T. (1996). Acer Quest: The Interactive Test Analysis System. In *Australian Council for Educational Research*.

Aiken, L. R. (1985). Three Coefficients For Analyzing The Reliability And Validity Of Ratings, Educational And Psychological Measurument. *Educational and Psychological Measurement*, *45*(1), 131–142.

Ayanwale, M. A., Isaac-oloniyo, F. O., & Abayomi, F. R. (2020). Dimensionality Assessment Of Binary Response Test Items : A Non-Parametric Approach Of Bayesian Item Response Theory Measurement. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(2). https://doi.org/10.11591/ijere.v9i2.20454

Brinkman, S. A., Kinnell, A., Maika, A., Hasan, A., Jung, H., & Pradhan, M. (2017). Validity and Reliability of the Early Development Instrument in Indonesia. *Child Indicators Research*, *10*(2), 331–352. https://doi.org/10.1007/s12187-016-9372-4

Crocker, L., & Algina, J. (2008). Introduction To Classical And Modern Test Theory-Procedures For Estimating Reliability. In *Harcourt Brace Jovanovich College*.

Eloy, A., Achutti, C. F., Fernandez, C., & Lopes, R. D. D. (2022). A Data-Driven Approach to Assess Computational Thinking Concepts Based on Learners' Artifacts. *Informatics in Education*, *21*(1), 33–54. https://doi.org/10.15388/infedu.2022.02

Gliem, J. A., & Gliem, R. R. (2003). The Preparation and Structure of Metalla-Sulphur/Selenium Nitrogen Complexes and Cages. *Research to Practice Conference in Adult, Continuing, and Community Education*, 82–88. https://doi.org/10.1016/B978-0-444-88933-1.50023-4

Herlina, K., Abdurrahman, A., & Umam, A. N. (2025). Expression-Based E-Worksheet (EBEW): An Effort to Enhace Student' Computational Thinking Skills. *Journal of Technology and Scince Education*, *15*(1), 50–63.

Hutchins, N. M., Biswas, G., Maróti, M., Lédeczi, Á., Grover, S., Wolf, R., Blair, K. P., Chin, D., Conlin, L., Basu, S., & McElhaney, K. (2020). C2STEM: a System for Synergistic Learning of Physics and Computational Thinking. *Journal of Science Education and Technology*, *29*(1), 83–100. https://doi.org/10.1007/s10956-019-09804-9

Kaya, D., Yaşar, A. Ö., Çetin, İ., & Kutluca, T. (2025). The Relationship Between The 21st-Century Skills And Computational Thinking Skills Of Prospective Mathematics And Science Teachers. *Journal of Pedagogical Research*, *9*(1), 73–95. https://doi.org/10.33902/JPR.202531498

Lai, R. P. Y. (2022). Beyond Programming: A Computer-Based Assessment of Computational Thinking Competency. *ACM Transactions on Computing Education*, *22*(2). https://doi.org/10.1145/3486598

Lu, C., Macdonald, R., Odell, B., Kokhan, V., Demmans Epp, C., & Cutumisu, M. (2022). A Scoping Review of Computational Thinking Assessments in Higher Education. *Journal of Computing in Higher Education*, *34*(2), 416–461. https://doi.org/10.1007/s12528-021-09305-y

Mariana, E. putri, & Kristanto, Y. dwi. (2023). Integrating STEAM Education and Computational Thinking: Analysis of Students' Critical and Creative Thinking Skills in an Innovative Teaching and Learning. *Southeast Asia Mathematics Education Journal*, *13*(1), 11–36.

Marina, R., Hapizah, H., & Hartono, Y. (2025). Validity of Student Worksheets Based on Computational Thinking With the Context of Local Wisdom in Musi Banyuasin to Support Numeracy Skills. *Mathline : Jurnal Matematika Dan Pendidikan Matematika*, *10*(2), 463–476. https://doi.org/10.31943/mathline.v10i2.940

Mukarom, Z., Darmawan, D., Agustin, M., Dwijantie, J. S., & Samadi, M. R. (2024). International Education Trend issues. *International Education Trend Issues*, *2*(2), 317–328.

Ocy, D. R., & Sarifah, I. (2025). EFA and CFA Analysis : Development and Validation of A Test Instrument For Mathematical Abstraction Skills. *Journal of Research and Advances in Mathematics Education*, *10*(2), 101–119.

Orban, C. M., & Teeling-Smith, R. M. (2020). Computational Thinking in Introductory Physics. *The Physics Teacher*, *58*(4), 247–251. https://doi.org/10.1119/1.5145470

Purnomo, H., Sa'dijah, C., Hidayanto, E., Sisworo, Permadi, H., & Anwar, L. (2022). Development of Instrument Numeracy Skills Test of Minimum Competency Assessment (MCA) in Indonesia. *International Journal of Instruction*, *15*(3), 635–648. https://doi.org/10.29333/iji.2022.15335a

Saritepeci, M., & Durak, A. (2022). Adaptation of T-STEM CT Scale to Turkish: Teacher Self-Efficacy and Outcome Expectancy for Teaching Computational Thinking. *Research on Education and Psychology*, *6*(Special Issue), 47–56. https://doi.org/10.54535/rep.1080132

Susanto, Kristiana, A. I., Dafik, Fatahillah, A., Venkatachalam, M., & Tanna, D. (2025).

Enhancing Students' Conjecturing Skills Through RBL-STEM with Antimagic Coloring and Geometric Transformation in Batik Design. *Emerging Science Journal*, *9*(2), 868–889. https://doi.org/10.28991/ESJ-2025-09-02-019

Tolba, E. G. (2025). Computational Thinking Skills and Computational Self-Efficacy as Predictors of Academic Well-Being. *International Journal of Technology in Education*, *8*(3), 761–788. https://doi.org/10.46328/ijte.1180

Version, D., & Dongen, V. (2020). Personality Disorders : Theory , Research , and Treatment. *Personality Disorders: Theory, Research, and Treatment*.

Yaniawati, P., Maat, S. M., Supianti, I. I., & Fisher, D. (2022). European Journal of Educational Research. *European Journal of Educational Research*, *11*(1), 69–81. https://pdf.eu-jer.com/EU-JER_9_1_395.pdf

Yin, Y., Hadad, R., Tang, X., & Lin, Q. (2020). Improving and Assessing Computational Thinking in Maker Activities: the Integration with Physics and Engineering Learning. *Journal of Science Education and Technology*, *29*(2), 189–214. https://doi.org/10.1007/s10956-019-09794-8

Zakwandi, R., & Istiyono, E. (2023). A framework for assessing computational thinking skills in the physics classroom: study on cognitive test development. *SN Social Sciences*, *3*(3), 46. https://doi.org/10.1007/s43545-023-00633-7

Zapata-Caceres, M., Martin-Barroso, E., & Roman-Gonzalez, M. (2021). Collaborative Game-Based Environment and Assessment Tool for Learning Computational Thinking in Primary School: A Case Study. *IEEE Transactions on Learning Technologies*, *14*(5), 576–589. https://doi.org/10.1109/TLT.2021.3111108

ZSOLDOS-MARCHIȘ, I., & BÁLINT-SVELLA, É. (2023). Development and preliminary testing of the AlgoPaint Unplugged Computational Thinking Assessment for Preschool Education. *Acta Didactica Napocensia*, *16*(1), 32–50. https://doi.org/10.24193/adn.16.1.3