Volume 11 Number 1, April 2026, pp. 84~96
P-ISSN:2477-7935
E-ISSN: 2548-6225
DOI: 10.59052/edufisika.v11i1.53979

DEVELOPMENT OF PHYSICS ASSESSMENT USING MARBLES GAME CONTEXT TO MEASURE CRITICAL THINKING SKILLS

Heldalia Heldalia[1], Yusman Wiyatmo [1]
[1]Physics Education, Universitas Negeri Yogyakarta, Daerah Istimewa Yogyakarta, Indonesia
Corresponding author email: heldalia.2024@student.uny.ac.id

Article Info
Received: 17 March 2026
Accepted: 02 April 2026
Publication: 09 April 2026

Abstract :
Critical thinking is a core competency in 21st-century education. However, students' critical thinking skills in physics learning remain relatively low, particularly in problem analysis and argument construction. Previous studies have developed various assessment instruments to measure critical thinking skills, but these instruments are generally not contextualized within local cultural practices and are rarely designed specifically for momentum and impulse topics. Furthermore, the integration of traditional games as a context for authentic assessment has not been widely explored. This study employed a Research and Development (R&D) design based on the ADDIE model, involving 324 eleventh-grade students from several schools in Yogyakarta for empirical validation. This instrument consists of eight essay items designed based on critical thinking indicators and validated by six validators. Content validity was analyzed using Aiken's V, while empirical validity, reliability, and item difficulty were examined using the Rasch model with the Partial Credit Model (PCM). The results show that all items achieved high content validity (Aiken's V = 0.833–1.00) and met the Rasch model fit criteria, as indicated by acceptable INFIT and OUTFIT MNSQ values and positive point measure correlations (PTMEA CORR). The instrument also demonstrated high reliability (Item Reliability = 0.87; Person Reliability = 0.85), with a moderate level of item difficulty. The instrument demonstrates strong validity and reliability, as well as high measurement precision in differentiating students' levels of critical thinking skills.

Keywords: Critical Thinking Skills, Traditional Marbles Game, Physics Assessment

INTRODUCTION

Critical thinking is a critical competence in the 6C skills framework and one of the main objectives of 21-century education because it fosters academic achievement and prepares for future professions (Anugrah & Astriani, 2024; Ariani, 2020; Asyhar, 2023; Sumanik, 2022). Such a life skill is vital in physics education as it empowers students to explain and make sense of phenomena based on scientific principles(Nurjanah et al., 2022; Sebastian & Kuswanto, 2025). However, despite

the great emphasis on such skills, students' ability to think critically is very poor in Indonesia, especially in terms of argument and problem solving. (Ariani, 2020; Husnah, 2018; Septiani et al., 2020).

At the same time, this issue is connected with the problem of limited quality of assessment practices that are dominated by recall tasks and have not been designed to effectively measure higher-order thinking skills. Prior studies indicate that creating valid and reliable instruments to measure students' critical thinking is a challenges at the secondary education level because of the complexity of the learning context (Anggiasari et al., 2018; Socrates & Mufit, 2022). As a result, current evaluative practices do not yet adequately provide for measuring students' critical thinking skills holistically.

Empirical findings from initial observations and interviews conducted at several public high schools (SMA) in the Special Region of Yogyakarta, namely SMAN 1 Sewon, SMAN 3 Bantul, and SMAN 1 Sedayu further underscore this issue. The results indicate that students' critical thinking skills are still categorized as low. Students experience significant difficulties in analyzing problem situations, identifying relevant physical principles, evaluating solution strategies, and constructing logical arguments when solving physics problems. Most students tend to rely on direct formula substitution without critically examining the conceptual basis of the problem. These findings reveal a discrepancy between the expected competency standards and the actual cognitive performance demonstrated by students in classroom assessments.

While existing previous research has generated instruments for measuring critical thinking skills in physics, these instruments are typically focused on a small subset of topics and do not contain the context necessary to make them relevant (Priyadi et al., 2022; Wahyu & Rukiyati, 2022), are not easily applicable to complex concepts like momentum and impulse, nor aligned with specific indicators of critical thinking. As a positive learning experience, local wisdom toward the use of traditional games has been acknowledged as a promising integration to bridge students' cultural context and real-life experiences into classroom learning(Sari et al., 2019; Wiyono et al., 2020). Yet, ingrained games like marbles that are becoming less frequent in students' experiences have not been extensively leveraged as an assessment context (Ansumarwaty & Busyairi, 2021; Ngaisah et al., 2023). his highlights that a cohesive approach to culture, subject-specificity and critical thinking has not yet emerged among current assessment instruments.

Momentum and impulse are topics generally considered within the domain of mechanics that often require deep conceptual understanding as learners need to interpret how dynamic interactions occur, analyze the events taking place during collision, and apply the law of conservation of momentum through multiple forms (mathematical, graphical, and conceptual). Force and impulse has also become difficult material to be differentiated by students, collision duration factor is a very important aspect in momentum changes (Himawan & Ariswan, 2023; Nabilah et al., 2020; Savira et al., 2019). Nonetheless, both current evaluation methods on these subjects focus more on performing numerical operations compared to writing logical proofs and arguments.

As concluded in the previous section, there is a dearth in past scholarship that investigates culturally contextualized assessment of topic-specific physics content (namely momentum and impulse) using universally clarified indicators for critical thinking in a singular psychometrically validated instrument. There is an evident need for developing a holistic physics assessment tool that is both topic-specific and aligned to critical thinking indicators as well as context specific, through culturally meaningful activities like traditional game of marbles. To solve this gap, an integrated assessment tool on momentum and impulse was developed which involved a classical marbles game to give a more validity usage on the authentic measurement of students critical thinking skills.

RESEARCH METHOD

Research Design

This is a research and development (R&D) study that aims to develop a physics assessment tool based on the conventional marble game to measure critical thinking skills of high school students. The ADDIE model (Branch, 2009): this development model is built around five main stages: Analysis, Design, Development, Implementation and Evaluation. We decided using the ADDIE model to identify instructional needs due to its systematic and whose proven method to develop valid, reliable and practical learning products and assessment instruments.

Research Target/Subject

The subjects in this research were 324 students of class XI from three public high schools in Yogyakarta namely SMAN 1 Sewon, SMAN 3 Bantul, and SMAN 1 Sedayu (by using purposive sampling). The schools were selected based on criteria, including that (1) the school has implemented the Merdeka Curriculum, (2) adequate physics learning facilities are available in the school, and (3) the willingness of participating in research. Additionally, students were selected based on their enrollment in physics classes which had already addressed the topics of momentum and impulse so that there would be a match between the content of the assessment instrument and students' previous learning experiences.

Furthermore, with variations in students' physics learning achievement levels across each school, a fairly heterogeneous sample was obtained from these research participants. This breadth of variability was desired to ensure the resultant tool works well for instruments at different ranges of student proficiency. Sample size (n = 324) was considered sufficient for Rasch model analysis, as it exceeded the number of respondents recommended in literature to achieve stable item parameter estimates and high test-retest reliability.

To obtain stable calibrations of items with a small standard error, a sample size of 324 individuals was selected, influenced by the statistical reasoning that samples over 250 provide highly stable estimates for parameters in Rasch models. Furthermore, expert validators were also invited to assess the content validity and construct validity of the developed assessment instrument which consists of 2 physics education lecturers and 4 competent practitioners.

Research Procedure

Analyze

Results: Critical thinking in learning physics was found to be needed based on the analysis stage This analysis incorporated a review of commonly-used in-school assessments, student characteristics and the potential for integrating the traditional marbles game context into momentum and impulse content. Based on the findings of this analysis, there was a need to develop contextual, authentic and culturally relevant tools.

Design

The next step is the design phase producing an initial version of the instrument that assesses. Activities include conducting assessments, preparing a blueprint based on critical thinking skills indicators (analyzing facts from the problem, formulating the main problem, providing logical arguments, and drawing conclusions), developing context-based test items around games using marbles material in physics subjects, and building an assessment rubric that follows the traits of critical thinking and physics concepts measured.

Eight essay questions were created and included all indicators of critical thinking skills. As a first step, these were clustered into 2 sets, Set A & Set B, both with 4 questions and it was ideally to address the distribution of indicators and the difficulty level of questions in the design phase However, in the implementation phase, all questions were collapsed into one single assessment instrument and it was firstly administered to all students. The instrument was applied in a pilot test to 324 students of three schools with different ability levels (i.e. high, moderate and low) in the eleven grade. By increasing the diversity in recovery, we support improved estimation of both item parameters and student ability within the Rasch model analysis. The instrument was developed using systematic alignment across physics content, critical thinking indicators, and cognitive levels as shown in table 1 below.

Table 1. Blueprint of critical thinking skills assessment instrument

Topic	Critical Thinking Skill Indicators	Cognitive Level	Item Number
Momentum and Impulse Theorem	Providing logical arguments	C5	1
	Analyzing facts from problems	C4	2
	Formulating key problem points	C4	3
	Drawing conclusions	C5	4

Conservation of Momentum and Collisions	Providing logical arguments	C5	5
	Analyzing facts from problems	C4	6
	Drawing conclusions	C5	7
	Formulating key problem points	C5	8

Development

To validate the draft instrument during development, two specialists in physics education and four experienced physics teachers served as validators. Validation was done to assess the instrument content, construct, and language. Questionnaire adjustments also reflected qualitative feedback on the instrument provided by validators. Validators' comments more generally addressed the question wording clarity, appropriateness of the marbles game context, alignment with indicators of critical thinking and accuracy of physics concepts. Such feedback formed the basis for modifying the instrument before moving onto Step Four.

Implementation

The implementation phase consisted of the pilot test of the revised instrument within students, while conducting learning activities based on marble-game. In this stage, the data for assessment were taken to assess the internal empirical validity, reliability and difficulty level of each item in the test. Rasch Model: The empirical validity analysis has been carried out by using the Winsteps application. Since the instrument contains items with multi-level scoring that requires an affirmativel answer, PCM is appropriate to analyze the data.

Evaluation

It does this by conducting formative and summative evaluations. Formative evaluation is conducted at each phase of the ADDIE model to enable improved learning materials continually. The analysis involves assessment of the identified needs for appropriateness and adequacy, as well as examination of the alignment that exists between curriculum specifications and critical thinking indicators. During the design stage, evaluation is performed by checking whether the question matrix, item specifications and assessment rubrics are in accordance with indicators and physics concepts with test items.

Evaluation during the development stage is performed by experts in practitioners and physics education faculty. The aspects valuable to assess content validity, construct validity, language clarity and further item relevance. Based on the feedback from experts, adjustments are made to develop a validated instrument. Field testing is an aspect of formative evaluation at the implementation stage. The quality of the instrument was analyzed with Rasch model including item fit statistics (infit and outfit mean square), Item difficulty test retests reliability, and separation index. Those items that did not pass were revised or removed.

After the implementation stage, a summative evaluation was completed to provide final feedback on instrument acceptability. The eligibility criteria for the instrument included: (1) good content and construct validity based on expert judgment, (2) item fit values ranging from 0.5-1.5, (3) person reliability ≥ 0.70, and (4) Good Discrimination Index that could differentiate students' ability level.

Data Analysis Technique

To assess the content validity of the instrument that has been developed, an expert validator conducted an evaluation using predefined criteria. Validators rated each item using a 4-point scale based on the item's suitability. The rating criteria were established as follows: (1) the item cannot be used, (2) the item can be used with extensive revisions, (3) the item can be used with minor revisions, and (4) the item can be used without revisions. The use of this 4-point scale aims to avoid neutral options so that validators provide more definitive ratings. The data obtained were analyzed by tabulating all validator scores and calculating the content validity coefficient using Aiken's V formula:

$$V = \frac{\sum s}{n(c-1)}$$

Explanation: $s = r - lo$, where lo is the lowest validation score and c is the highest validation score. The results of the analysis of Aiken's v validity index are categorized according to Table 2 of the following instrument validity criteria.

Table 2. Instrument validity criteria

No.	Score Range	Category
1	$0{,}8 < V \leq 1$	Very High
2	$0{,}6 < V \leq 0{,}8$	High
3	$0{,}4 < V \leq 0{,}6$	Moderate
4	$0{,}2 < V \leq 0{,}4$	Low

Item fit was analyzed using several Rasch parameters: Infit Mean Square (INFIT MNSQ), Outfit Mean Square (OUTFIT MNSQ), and Point Measure Correlation (PTMEA CORR). INFIT MNSQ is used to detect response discrepancies on items corresponding to students' ability levels, while OUTFIT MNSQ is sensitive to extreme responses on items that are too easy or too difficult. PTMEA CORR indicates the correlation between item scores and overall ability, thereby reflecting the item's consistency with the construct being measured.

The range of INFIT MNSQ values used is 0.77–1.33, indicating that the item is productive in measurement and capable of distinguishing student ability effectively (Bond & Fox, 2015). Meanwhile, acceptable OUTFIT MNSQ values fall within the range of 0.5–1.5, and a PTMEA CORR value ≥ 0.30 indicates that the item has a positive correlation with the construct being measured. The INFIT Mean of Square value was used to determine whether an item passed or failed based on Table 3.

Table 3. INFIT mean of square value category

Mean square INFIT Value	Category
> 1,33	Very Suitable
0,77– 1,33	Suitable
< 0,7	Not Suitable

Then, the difficulty level of each item is analyzed by looking at the delta value (b). The assessment categories are as follows in Table 4 (Setyawarno, 2017).

Table 4. Difficulty level assessment categories

Score	Category
$b > 2$	Very Difficult
$1 < b \leq 2$	Difficult
$-1 < b \leq 1$	Moderate
$-1 < b \leq -2$	Easy
$b < -2$	Very Easy

Reliability analysis was conducted using the Winsteps software based on the Rasch model. In Rasch modeling, reliability is divided into two categories: item reliability and person reliability. Item reliability indicates the consistency of item difficulty estimates across different samples, reflecting the extent to which test items form a stable measurement scale. A high item reliability score indicates that the distribution of item difficulty is appropriate and effectively represents the construct being measured. Conversely, person reliability indicates the consistency of students' performance in answering test items, reflecting the instrument's ability to distinguish students with different ability levels. A high person reliability value indicates that the instrument is sensitive to variations in students' critical thinking skills.

Reliability values were obtained from the summary of item estimates and the summary of person (case) estimates. The criterion used was a reliability value of ≥ 0.70, indicating that the instrument is considered reliable. The reliability value categories are presented in Table 5.

Table 5. Value category reliability level

Reliability Value (R)	Reliability Level
0,80 ≤ R ≤ 1,00	High
0,60 ≤ R < 0,80	Moderately High
0,40 ≤ R < 0,60	Moderate
0,20 ≤ R < 0,40	Low
0,00 ≤ R< 0,20	Very Low

Furthermore, the reliability results obtained demonstrate the instrument's ability to measure variations in student performance. High item reliability indicates that the instrument has a stable hierarchy of difficulty levels, while high test-taker reliability indicates that the instrument is effective in distinguishing students' levels of critical thinking ability. Thus, both indices support the instrument's quality as a reliable measurement tool.

RESULTS AND DISCUSSION

Results

In determining the feasibility of the instrument, an evaluation was conducted focusing on three criteria, namely construct, content, and language aspects. The data from the feasibility assessment was then processed using Aiken's V method to determine the validity level of each item that had been developed.

Table 6. Results of the critical thinking skills instrument feasibility test

Question Item	*Aiken's V*	Category
A3, B2, B3	1	high Aiken's V values
A2, B4	0,944	high Aiken's V values
A1, B1	0,888	high Aiken's V values
A4	0,833	high Aiken's V values

Based on the results of the data analysis in the table, each item in the instrument was found to have an Aiken's V value indicating a high level of fit. The range of values obtained was between 0.833 as the minimum limit and 1 as the maximum value. With these consistent results, all eight items, covering codes A1 to B4, were classified as highly feasible. These results indicate strong evidence of content validity, supporting the instrument's suitability for data collection.

To verify that each item functions properly, the data was analyzed using Winsteps software based on the Rasch model with the Partial Credit Model (PCM). This analysis evaluates several item characteristics, including item fit, difficulty level, and reliability. Item fit is assessed using three primary indices: INFIT MNSQ, OUTFIT MNSQ, and Point Measure Correlation (PTMEA CORR). An item is considered to fit the Rasch model if its INFIT and OUTFIT MNSQ values fall within the acceptable range of 0.77 to 1.30, while a positive PTMEA CORR value indicates that the item is consistent with the overall construct being measured. The results of item fit statistics of critical thinking skills instrument based on rasch model analysis can be seen in Table 7.

Table 7. Item Fit Statistics of Critical Thinking Skills Instrument Based on Rasch Model Analysis

Question	INFIT MNSQ	OUTFIT MNSQ	PTMEA CORR	Decision
1	1,30	1,33	0,64	Fit
2	1,02	1,02	0,67	Fit
3	1,06	1,06	0,65	Fit
4	1,11	1,06	0,65	Fit
5	0,94	0,84	0,67	Fit
6	0,94	0,87	0,72	Fit
7	0,80	0,74	0,70	Fit

8	0,92	0,89	0,61	Fit

Based on the analysis results presented in Table 7, all items met the established fit criteria. The INFIT MNSQ values ranged from 0.80 to 1.30, with OUTFIT MNSQ values ranging from 0.74 to 1.33, and the PTMEA CORR is positive, the item is considered acceptable for measurement. These results confirm that all items align with the Partial Credit Model (PCM) and are suitable for further use in measuring the intended construct.

When compared to previous studies, the results of this study remain within a normal and acceptable range. A number of Rasch-based studies report that the presence of a few misfit items is common in instrument development and often requires revision to improve construct validity (Ayoub et al., 2024; Fährmann et al., 2022). However, there are also studies indicating that all items can meet fit criteria without significant misfits, provided the items are formulated consistently with the construct being measured. This suggests that the "all items fit" finding in this study remains consistent with contemporary Rasch research practices.

Nevertheless, the condition where all items demonstrate good fit must be interpreted with caution. The absence of item misfit may indicate sample homogeneity or low response variation, which could potentially reduce the instrument's sensitivity in distinguishing students' ability levels. Therefore, in addition to item fit analysis, it is important to consider other Rasch indicators, such as item difficulty distribution and discrimination indices, to ensure that the instrument possesses adequate discriminatory power.

The reliability of the critical thinking skills test items is based on the person reliability and item reliability values as output from the Winsteps software. The detailed results of this analysis are presented in Table 8 below.

Table 8. Reliability values of critical thinking skills instruments

Output	Reliability Value	Category
Item Reliability	0,87	Very High Reliability
Person Reliability	0,85	Very High Reliability

Based on the results of data processing using the Rasch model, this instrument has been proven to have an excellent level of consistency. This can be seen from the Item Reliability value of 0,87 and Person Reliability of 0,85, both of which fall into the Highly Reliable category.

The high Item Reliability value indicates that the quality of the test items is very stable and of high quality. Meanwhile, the Person Reliability value confirms that the response patterns of the students are very consistent, so that the test results can be trusted to differentiate their skill levels. Overall, these figures assure that the instrument indicates acceptable consistency and is suitable for use as a research data collection tool.

The level of difficulty of the questions on the critical thinking skills test is determined based on the item measure value generated as output from the Winsteps software. The criteria used stipulate that an item has a good level of difficulty if the difficulty value is in the range of -2.00 to 2.00. The difficulty level of the critical thinking test instrument is presented in Table 9.

Table 9. Results of item difficulty level analysis

Aspect	Question Item	Measure	Category
Critical Thinking Skills	1	-0,57	Moderate
	2	+0,21	Moderate
	3	+0,18	Moderate
	4	+0,12	Moderate
	5	-0,04	Moderate
	6	-0,18	Moderate
	7	-0,04	Moderate
	8	+0,32	Moderate

Table 9 shows that in critical thinking skills, item number 1 has the lowest difficulty value of -0.57. This value is close to 1, which means that the item is in the moderate category but close to the easy category. In critical thinking skills, item number 8 has the highest difficulty value of +0.32. The compatibility of critical thinking and verbal representation items with the Partial Credit Model (PCM) can also be observed in Figure 1.

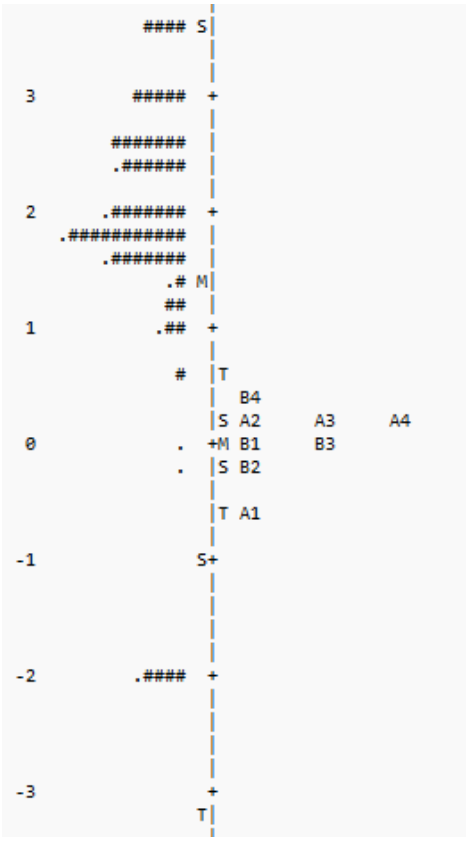

Figure 1. Wright Map of Item Difficulty and Person Ability Distribution

Figure 1 shows that the determination of item difficulty levels in the critical thinking test was carried out to ensure that the variation in question difficulty levels matched the spectrum of student abilities. Items with a moderate level of difficulty were expected to clearly differentiate between students with good and low conceptual understanding, thus ensuring that the instrument was well calibrated. Specifically, item B4 was identified as the most difficult item, while item A1 was the easiest. This variation indicates that the instrument covered the range of difficulty necessary to reflect differences in student competency.

However, based on the Wright map, it can be seen that the average student ability (mean person) was above the average item difficulty level (mean item), indicating that the instrument tended to be relatively easy for the sample studied. This condition indicates that the majority of students had abilities above the average item difficulty level; the instrument's ability to differentiate high-ability students was less than optimal. Furthermore, there were indications of a distribution gap, particularly at the great difficulty level, where the number of items targeting high-ability students was still limited. This shows that although the instrument has functioned well in general, further development is needed by adding more challenging items so that the measurement coverage is more even and the measurement precision is increased.

In addition to quantitative findings, the instrument development process in this study was also enriched by qualitative feedback from expert validators. The feedback provided primarily concerned changing the mass of the marbles from 15 grams to 5 grams, improving the correct formula writing, revising the rubric to be clearer and more detailed, changing the marble's fall height from 40 cm to 15 cm, and improving the placement of images.

Discussion

In terms of feasibility, we assessed construct, content and language. Aiken's V was used to explore content validity, determining the level of expertise agreement for appropriateness of each item according to the targeted critical thinking indicators.

As shown in the analysis, all items reached high Aiken's V values (from 0.833 to 1.000). This pattern indicates consensus among validators, indicating that the items sufficiently represent the intended facets of critical thinking. In the other hand, it means that at the item level, the construct has been operationalized into measurable indicators in a consistent manner. Although previous studies Aristiawan & Istiyono, 2020 and Astuti et al (2020) report akin findings yet this study locates these indicators in a more specific and cognitively demanding context of momentum and impulse making the construct more interpretable.

To explore empirical functionality of these items, a Rasch analysis was performed using the Partial Credit Model (PCM) in Winsteps. Performance of item fit was assessed through multiple indices such as INFIT MNSQ, OUTFIT MNSQ and Point Measure Correlation (PTMEA CORR) All items fell within a satisfactory range as demonstrated by the results. These results showed that the items worked well consistently in the model structures and with respect to the construct. No misfitting items were found, indicating the measurement structure is stable and coherent.

The results suggest that the instrument operates around one common latent variable, which is not strongly distorted with irrelevant variance. There are no discrepant items so each item contributes equally to the measuring. In comparison to some previous studies practicing generic contexts (Naqiyah et al., 2020; Permatasari et al., 2019), these findings indicate that it is possible to achieve steady internal consistency in a domain-specific and context-based manner. However, uniformity in concordance across items should not be taken at face value: it can also occur due to restricted variance in responses or durable concordance induced by familiarity with the context of measures rather than exclusively with that of a construct targeted.

A more fundamental issue is whether the context of the marbles game in fact capture processes of critical thinking or only incompletely reflects those already familiar to students with the scenario. If students' performance will depend on how well they recognize and understand the context, then the instrument risks conflating contextual familiarity with critical thinking skills. This leads to an increase in construct-irrelevant variance due to irrelevant factors influencing the measurement of that construct. Moreover, the implementation of a single/culture context could add biases to students with lower exposure to common marble games who might not effectively interpret this item. While the Rasch model provides statistical fit and internal consistency, it does not solve construct validity challenges associated with context dependency. Hence, other investigation (especially through cross-group validation and in different contexts) remains necessary to demonstrate that the instrument captures critical thinking skills rather than familiarity with the context.

Also, the reliability analysis sustains the quality of the instrument. The item reliability (0,87) and person reliability 0,85 reflects a strong level of constancy. These values reflect the stability of the instrument on item hierarchy and the consistency in capturing students' response patterns rather than reflecting mere statistical adequacy. This is adequate, because it means that the instrument can detect differences in students' critical thinking ability with sufficient precision.

With respect to item difficulty, all items range around moderate difficulty (−0.57 to +0.32 logits). The 1st item seems comparatively easier whereas the 8th is the most complex of all. But none are in the extremes. This balanced distribution is critical because it enables the instrument to detect variation in students' abilities without allowing the items to be pejoratively skewed toward either too easy or too difficult.

From a measurement perspective, the fact that they are able to target this type of item suggests that the instrument is appropriately matched to their ability level. The distribution also prevents ceiling and floor effects, so the instrument is sensitive at all levels of performance. So we can see differences not only in correct answers but also in what makes for better reasoning behind them.

Looking at these findings from a theoretical perspective, they reinforce the idea that critical thinking is central to physics learning, especially in topics like momentum and impulse. These topics require more than applying formulas; they involve interpreting situations, analyzing relationships, and making reasoned judgments (Hapsari, 2016; Ufairiah & Laksanawati, 2020). As described by Facione

(2023) and (Ennis, 2015), critical thinking includes interpretation, evaluation, and reflective decision-making. The current instrument attempts to capture these processes in a structured way, allowing assessment to move beyond procedural knowledge toward deeper conceptual understanding (Ritdamaya & Suhandi, 2016; Sundari & Sarkity, 2021).

Another aspect that strengthens this study is the use of a cultural context through traditional marbles game. From a contextual and sociocultural perspective, learning becomes more meaningful when it is connected to familiar experiences (Annisha, 2024; Asra et al., 2021). By embedding such context into the items, the instrument does not only measure thinking skills but also engages students in a more relatable way (Pratiwi & Pujiastuti, 2020). Compared to conventional assessments that often rely on abstract situations, this approach brings the problems closer to students' everyday experiences, which can support deeper cognitive engagement.

The contribution of this study lies in the explicit integration of three components that have rarely been operationalized simultaneously in previous research. Previous studies on critical thinking assessment in physics have generally focused on measuring students' abilities using Rasch analysis without specifically integrating critical thinking indicators into item levels (Afandi et al., 2025; Kassiavera et al., 2024). Additionally, some studies have developed instruments on specific physics topics, such as fluid dynamics or climate change, but remain limited to content-based measurement without simultaneously incorporating critical thinking indicators and context (Santoso & Wuryandani, 2020). In contrast to these studies, this study translates critical thinking skills into specific and measurable indicators that are directly integrated into each test item, focusing on momentum and impulse as conceptually challenging topics.

Furthermore, although some recent studies have begun to integrate local wisdom contexts into learning and assessment, such contexts generally serve only as supplementary elements and have not yet become structural components in item construction (Annisha, 2024; Asra et al., 2021). From a methodological perspective, this study offers a systematic approach to developing context-based physics assessment instruments that simultaneously integrate critical thinking indicators, specific content, and cultural context within the framework of the Rasch Partial Credit Model. From a practical standpoint, the resulting instruments are not only psychometrically validated but also contextual and applicable, allowing teachers to use them directly to measure students' critical thinking skills in physics learning.

That said, some limitations need to be considered. The study was conducted in a limited number of schools within one region, which means the findings should be interpreted with caution when applied to broader contexts. In addition, the instrument focuses on a specific topic, and further studies are needed to explore its use in other areas of physics. Future research could also examine how similar approaches work in different cultural or educational settings.

CONCLUSION

This research was a development of the physics assessment tool intensity and impulse were integrated with a traditional marbles's game to measure high school students' critical thinking skills. These findings demonstrate that this assessment tool meets solid psychometric criteria. The content validity of all items is high according to Aiken's V, the Partial Credit Model meets the Rasch model fit requirements and has high item and individual reliability indices. The moderately difficult item distribution here signifies that the measuring gadget is well adjusted to separate pupils within the varied vital thinking degree. Findings confirm that explicit incorporation of these critical thinking indicators within a culturally relevant context yields meaninfully substantial and statistically robust measurement instrument. The transitional situation of the game context helps students make connections between real experiences and more abstract concepts such as momentum and impulse and so is a necessity for participating in higher order thinking processes such that require analysis, evaluation, inference etc., rather than just applying formulas procedurally. Thus, this tool may be appropriate for classroom assessment and educational research to assess higher-order thinking skills in the discipline of physics. However, there are limitations to this study as it used a relatively homogeneous sample of high school students and concentrated only on the topics of momentum and impulse meaning that the findings may not be generalizable to other student populations or other physics concepts. This tool can be extended in follow-up studies on a larger population and its potential long-term effects on the enhancement of students' physics-related critical thinking skills should also be explored, as well as studying a bigger

sample including students who study different areas of knowledge and evaluate other topics like dynamics to broaden your extension.

ACKNOWLEDGMENTS

The author would like to thank Yogyakarta State University for its support and guidance in conducting this research and preparing this article. In addition, the author would also like to thank LPDP for the opportunity to conduct this research.

REFERENCES

Afandi, A., Aviyanti, L., Saepuzaman, D., & Setiawan, A. (2025). Assesing Critical Thinking Skills in Dynamic Fluids at Senior High School : Rasch Model Analysis. 14(2), 87–100. https://doi.org/10.24042/jipfalbiruni.v14i2.27836

Anggiasari, T., Hidayat, S., & Harfian, B. A. A. (2018). Analisis Keterampilan Berpikir Kritis Siswa Sma Di Kecamatan Kalidoni Dan Ilir Timur Ii. Bioma : Jurnal Ilmiah Biologi, 7(2), 183–195. https://doi.org/10.26877/bioma.v7i2.2859

Annisha, D. (2024). Integrasi Penggunaan Kearifan Lokal (Local Wisdom) dalam Proses Pembelajaran pada Konsep Kurikulum Merdeka Belajar. Jurnal Basicedu, 8(3), 2108–2115.

Ansumarwaty, F., & Busyairi, A. (2021). Analisis Hukum Kekekalan Momentum pada Permainan Tradisional Kelereng dengan Menggunakan Video Stop Motion untuk Meningkatkan Motivasi Belajar Peserta Didik. Jurnal Ilmiah Profesi Pendidikan, 6(3), 517–521. https://doi.org/10.29303/jipp.v6i3.243

Anugrah, J. I., & Astriani, D. (2024). Meningkatkan Keterampilan Berpikir Kritis Menggunakan Model Problem Based Learning Berbasis Literasi Sains. Pensa E-Jurnal: Pendidikan Sains, 12(2), 38–42.

Ariani, T. (2020). Analysis of Students' Critical Thinking Skills in Physics Problems. Kasuari: Physics Education Journal (KPEJ), 3(1), 1–17. https://doi.org/10.37891/kpej.v3i1.119

Aristiawan, A., & Istiyono, E. (2020). Developing Instrument of Essay Test to Measure the Problem-Solving Skill in Physics. Jurnal Pendidikan Fisika Indonesia, 16(2), 72–82. https://doi.org/10.15294/jpfi.v16i2.24249

Asra, A., Festiyed, F., Mufit, F., & Asrizal, A. (2021). Pembelajaran Fisika Mengintegrasikan Etnosains Permainan Tradisional. Konstan - Jurnal Fisika Dan Pendidikan Fisika, 6(2), 66–73. https://doi.org/10.20414/konstan.v6i2.67

Astuti, A. T., Supahar, Mundilarto, & Istiyono, E. (2020). Development of assessment instruments to measure problem solving skills in senior high school. Journal of Physics: Conference Series, 1440(1). https://doi.org/10.1088/1742-6596/1440/1/012063

Asyhar, B. (2023). Analysis of the Inquiry-Infusion Learning Model to Develop Students ' Critical Thinking Ability. 6(1), 1–18. https://doi.org/10.30762/f

Ayoub, A. E. A., Aljughaiman, A. M., Alghawi, M. A., Morsy, A., Omara, E. M. N., Abdulla Alabbasi, A. M., & Renzulli, J. S. (2024). Validation of Hamdan intelligence scale in upper elementary grades using the Rasch model: exploratory study. Frontiers in Psychology, 15(August), 1–11. https://doi.org/10.3389/fpsyg.2024.1407734

Ennis, R. (2015). What Is Critical Thinking in Higher Education? The Palgrave Handbook of Critical Thinking in Higher Education, 27–29. https://doi.org/10.1057/9781137378057.0004

Facione, P. A. (2023). Critical Thinking : What It Is and Why It Counts 2023 Update. Insight Assessment, ISBN 13: 978-1-891557-07-1., 1–28.

Fährmann, K., Köhler, C., Hartig, J., & Heine, J. H. (2022). Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies. Large-Scale Assessments in Education, 10(1). https://doi.org/10.1186/s40536-022-00124-w

Hapsari, S. (2016). A Descriptive Study of the Critical Thinking Skills of Social Science at Junior High School. Journal of Education and Learning (EduLearn), 10(3), 228–234. https://doi.org/10.11591/edulearn.v10i3.3791

Himawan, N. A., & Ariswan. (2023). Physics Learning E-Module Integrated with Practicing Pancasila Values on Momentum and Impulse : Is it Effective to Improve Students ‟ Critical Thinking Skill and Hard Work Character ? JIPF (JURNAL ILMU PENDIDIKAN FISIKA), 8(1), 30–41.

Husnah, M. (2018). Hubungan Tingkat Berpikir Kritis Terhadap Hasil Belajar Fisika Siswa Dengan Menerapkan Model Pembelajaran Problem Based Learning. PASCAL (Journal of Physics and Science Learning), 1(2), 10–17. https://doi.org/10.30743/pascal.v1i2.338

Kassiavera, S., Suparmi, A., Cari, C., & Sukarmin, S. (2024). Application of Rasch Model in Two-Tier Test for Assessing Critical Thinking in Physics Education. Journal of Baltic Science Education, 23(6), 1227–1242. https://doi.org/10.33225/jbse/24.23.1227

Nabilah, M., Sitompul, S. S., & Hamdani, H. (2020). Analisis Kemampuan Kognitif Peserta Didik Dalam Menyelesaikan Soal Momentum Dan Impuls. Jurnal Inovasi Penelitian Dan Pembelajaran Fisika, 1(1), 1. https://doi.org/10.26418/jippf.v1i1.41876

Naqiyah, M., Rosana, D., Sukardiyono, & Ernasari. (2020). Developing instruments to measure physics problem solving ability and nationalism of high school student. International Journal of Instruction, 13(4), 921–936. https://doi.org/10.29333/iji.2020.13456a

Ngaisah, N. C., Ayyubi, M. Al, Wati, L. N., Fajzrina, Aulia, R., Munawarah, Fadillah, C. N., & Zohro, N. P. (2023). Permainan Tradisional Kelereng dan Perannya dalam Mengembangkan Keterampilan Sosial Anak Nur. Jurnal Ilmiah Potensia, 8(1), 103–113.

Nurjanah, S., Djudin, T., & Hamdani. (2022). Analisis Kemampuan Berpikir Kritis Peserta Didik pada Topik Fluida Dinamis. Jurnal Education and Development, 10(3), 111–116.

Permatasari, A. K., Istiyono, E., & Kuswanto, H. (2019). Developing Assessment Instrument to Measure Physics Problem Solving Skills for Mirror Topic. International Journal of Educational Research Review, 4(3), 358–366. https://doi.org/10.24331/ijere.573872

Pratiwi, J. W., & Pujiastuti, H. (2020). Eksplorasi Etnomatematika Pada Permainan Tradisional Kelereng. Jurnal Pendidikan Matematika Raflesia, 05(02), 1–12.

Priyadi, R., Mustajab, A., Tatsar, M. Z., & Kusairi, S. (2022). Analisis Kemampuan Berpikir Kritis Siswa SMA Kelas X MIPA dalam Pembelajaran Fisika. Jurnal Pendidikan Fisika Tadulako (JPFT), 6(1), 53–55.

Ritdamaya, D., & Suhandi, A. (2016). Konstruksi Instrumen Tes Keterampilan Berpikir Kritis Terkait Materi Suhu dan Kalor. JPPPF - Jurnal Penelitian & Pengembangan Pendidikan Fisika, 2(2), 87–96.

Santoso, R., & Wuryandani, W. (2020). Pengembangan Bahan Ajar PPKn Berbasis Kearifan Lokal Guna Meningkatkan Ketahanan Budaya Melalui Pemahaman Konsep Keberagaman. Jurnal Ketahanan Nasional, 26(2), 229. https://doi.org/10.22146/jkn.56926

Sari, F. P., Nikmah, S., Kuswanto, H., & Wardani, R. (2019). Developing Physics Comic Media a Local Wisdom: Sulamanda (Engklek) Traditional Game Chapter of Impulse and Momentum. Journal of Physics: Conference Series, 1397(1). https://doi.org/10.1088/1742-6596/1397/1/012013

Savira, Y. M., Budi, A. S., & Supriyati, Y. (2019). Pengembangan E-Modul Materi Momentum Dan Impuls Berbasis Process Oriented Guided Inquiry Learning (Pogil) Untuk Meningkatkan Kemampuan Berpikir. Prosiding Seminar Nasional Fisika (E-Journal), VIII, Jakarta: Universitas Negeri Jakarta. https://doi.org/10.21009/03.SNF2019

Sebastian, R., & Kuswanto, H. (2025). The effectiveness of a physics e-book on rotational dynamics of a traditional top game assisted by augmented reality to improve students' critical thinking skills and visual representations. Education in Physics, Revista Mexicana de F´ısica E22 020205, 1–12. https://doi.org/10.31349/RevMexFisE.22.020205

Septiani, D., Riyadi, & Triyanto. (2020). Students' Verbal Representation Abilities in Solving Geometry Problems. Journal of Physics: Conference Series, 1594(1). https://doi.org/10.1088/1742-6596/1594/1/012046

Setyawarno. (2017). Upaya peningkatan kualitas butir soal dengan analisis aplikasi Quest. Disampaikan dalam Workshop dalam rangka kegiatan pelatihan guru IPA SMP Sleman Yogyakarta.

Socrates, T. P., & Mufit, F. (2022). Efektivitas Penerapan Media Pembelajaran Fisika Berbasis Augmented Reality: Studi Literatur. EduFisika: Jurnal Pendidikan Fisika, 7(1), 96–101. https://doi.org/10.59052/edufisika.v7i1.19219

Sumanik, N. B. (2022). Pengembangan Lembar Kerja Peserta Didik Elektronik Berbasis Literasi Sains untuk Melatih Kemampuan Berpikir Kritis. Paedagogia, 25(2), 147. https://doi.org/10.20961/paedagogia.v25i2.64080

Sundari, P. D., & Sarkity, D. (2021). Keterampilan Berpikir Kritis Siswa SMA pada Materi Suhu dan Kalor dalam Pembelajaran Fisika. Journal of Natural Science and Integration, 4(2), 149–161.

Ufairiah, Q. R., & Laksanawati, W. D. (2020). Identifikasi Masalah Kemampuan Berpikir Kritis Siswa Guna Mengetahui Pengaruh Model Dan Pendekatan Pembelajaran. Jurnal UNSIQ, 2(1), 75–82.

Wahyu, H. A., & Rukiyati. (2022). Studi literatur: Permainan tradisional sebagai media alternatif stimulasi perkembangan anak usia dini. Jurnal Pendidikan Anak, 11(2), 109–120.

Wiyono, K., Ismet, I., & Saparini, S. (2020). Development of interactive multimedia for learning physics based on traditional games. National Conference on Mathematics Education (NaCoME). https://doi.org/10.1088/1742-6596/1480/1/012074