Physics Communication
Volume 10 Issue 1 Year 2026 Pages 7-16
P-ISSN 2528-5971 | E-ISSN 2528-598X
https://journal.unnes.ac.id/journals/pc
https://doi.org/10.15294/pc.v10i1.18527

Integrating Claim, Evidence, Reasoning into a Four-Tier Diagnostic Test to Assess Students' Conceptual Understanding of Work and Energy

Muhammad Fatikhul Alam Bima Sakti*, Sunyoto Eko Nugroho, Budi Naini Mindyarto

Master of Physics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

*Correspondence to: *bimasakti1960@students.unnes.ac.id*

Abstract: This study reports the development and validation of a Four-Tier Diagnostic Test integrated with the Claim, Evidence, and Reasoning (CER) framework to examine senior high school students' conceptual understanding of work and energy. Although four-tier diagnostics and CER have been widely implemented independently, limited research has combined both approaches within a unified diagnostic assessment framework. The integration is intended to capture students' answer accuracy, confidence level, and the structure of their scientific reasoning simultaneously. Using a Research and Development design, five CER-based four-tier items were developed and administered to 100 twelfth-grade students. Content and construct validity were established through expert review, and reliability was examined using Cronbach's Alpha. Misconceptions were analysed using the Confidence Discrimination Quotient (CDQ) and four-tier interpretation criteria. The findings indicate that misconception rates ranged from 23% to 25%, while CDQ values varied between −0.006 and 0.024, suggesting the presence of moderate conceptual uncertainty. Most items demonstrated high difficulty levels and moderate discrimination indices, indicating that work–energy concepts remain cognitively demanding. The integration of CER into the four-tier structure provided richer diagnostic information by linking correctness, reasoning quality, and confidence level. These results highlight the potential of CER-integrated diagnostic assessment in identifying nuanced misconceptions in physics learning.

Keywords: Four-Tier Diagnostic Test; CER (Claim, Evidence, Reasoning); Conceptual Understanding; Misconceptions; Work and Energy

Submitted: 2024-12-21. **Revised**: 2025-12-01. **Accepted**: 2026-02-10.

Introduction

Work and energy are fundamental concepts in physics that underpin broader principles such as conservation laws, mechanical systems, and energy transfer processes. A solid understanding of these concepts is essential for students because they form the basis for advanced topics in mechanics and other branches of physics (Eliza et al., 2022; Kalies, 2022). Despite their central role, research consistently reports that students experience persistent conceptual difficulties in this domain.

Studies indicate that students frequently confuse force with work, misinterpret negative work, and struggle to apply the principle of energy conservation in dynamic contexts (Sulistri & Lisdawati, 2017; Wilantika et al., 2018). Many learners rely on everyday interpretations of "work" rather than its scientific definition, leading to fragmented understanding. Furthermore, students often demonstrate procedural competence in solving equations while lacking conceptual clarity about the underlying physical principles (Sheftyawan et al., 2018). These issues highlight the need for assessment approaches capable of uncovering not only incorrect answers but also the reasoning patterns that sustain misconceptions.

Traditional multiple-choice assessments are limited in diagnosing misconceptions because they primarily measure answer accuracy without revealing students' confidence levels or reasoning structures. In response to this limitation, diagnostic assessment models have been developed. The Four-Tier Diagnostic Test (FTDT) extends conventional formats by incorporating confidence ratings for both answers and reasoning, enabling differentiation between lack of knowledge and firmly held misconceptions (Setiawan & Faoziyah, 2020; Widayati et al., 2020). Previous studies have demonstrated the utility of FTDT in identifying

misconceptions across physics topics; however, most implementations focus on selecting predefined reasons rather than explicitly structuring scientific argumentation.

The Claim–Evidence–Reasoning (CER) framework, widely used in science education, emphasizes the construction of scientific explanations through explicit articulation of claims supported by evidence and justified by reasoning. CER has proven effective in strengthening students' argumentation skills and conceptual coherence. Nevertheless, its application has predominantly been instructional rather than diagnostic. Limited studies have integrated CER systematically into multi-tier diagnostic formats to enhance conceptual diagnosis.

Although FTDT and CER have been investigated independently, there remains a gap in research examining their integration within a unified diagnostic instrument for work and energy concepts. Combining these approaches may provide richer insight into students' conceptual stability by simultaneously capturing answer correctness, confidence level, and the structure of scientific reasoning. Therefore, this study aims to develop and validate a CER-integrated Four-Tier Diagnostic Test to assess senior high school students' conceptual understanding and misconceptions related to work and energy. Specifically, this study examines the instrument's validity and reliability, analyzes its item characteristics, and explores its capacity to identify patterns of misconception in this conceptual domain.

Methods

This research employs a Research and Development (R&D) approach. The study was conducted with 12th-grade students, totaling 100 participants, at SMAN 34 Jakarta. This study employed four data collection techniques: (1) a four-tier diagnostic test as the primary data source, (2) semi-structured interviews for explanatory follow-up, (3) a questionnaire to capture students' perceptions and confidence patterns, and (4) documentation to support contextual interpretation. Google Forms was used as the platform for piloting and administering the main test and questionnaire.

Four-Tier Diagnostic Test Integrated with Claim–Evidence–Reasoning

The diagnostic instrument consisted of four response tiers: (i) selection of a conceptual answer, (ii) confidence level in the selected answer, (iii) selection of reasoning structured according to Claim–Evidence–Reasoning (CER), and (iv) confidence level in the selected reasoning. The instrument was first piloted using Google Forms to examine item clarity and response functionality, followed by revision based on pilot feedback. Subsequently, Google Forms was used to administer the main data collection and to compile responses in a standardized format. This technique generated quantitative and categorical data, including item-level correctness, confidence ratings for answers and reasoning, CER reasoning patterns, and classification of students' understanding profiles (scientific understanding, misconception, lack of understanding, or guessing). These data directly addressed the study's objective by diagnosing the level and structure of students' conceptual understanding of work and energy.

Semi-Structured Interviews

Semi-structured interviews were conducted with a purposive subset of students representing diverse test outcomes (e.g., correct–high confidence, incorrect high confidence, low confidence). The interviews elicited students' verbalized claims, supporting evidence, and reasoning processes underlying their test responses. Interview data produced qualitative explanations of students' conceptual reasoning and sources of difficulty. These data were used to clarify ambiguous response patterns, validate the interpretation of diagnostic results, and provide triangulation between selected responses and articulated reasoning, thereby strengthening the explanatory power of the test findings.

Questionnaire

A brief questionnaire was administered via Google Forms following the test. The questionnaire

captured students' self-reported perceptions of item difficulty, clarity of CER components, and general confidence in learning work and energy concepts. The questionnaire yielded descriptive data on perceived difficulty and confidence trends that complemented the diagnostic outcomes. These data supported the interpretation of response confidence patterns and provided contextual information on how students experienced the CER-integrated diagnostic format.

Documentation

Documentation included the finalized instrument blueprint, item specifications, and records of pilot revisions. These materials provided contextual evidence of the instrument development process and supported the transparency and traceability of data collection. Collectively, the four techniques produced complementary datasets diagnostic scores and profiles (test), explanatory reasoning (interviews), perception and confidence indicators (questionnaire), and development records (documentation) that together enabled a comprehensive assessment of students' conceptual understanding of work and energy.

The expert validation questionnaire uses a Likert scale. According to Sugiyono (2017), the Likert scale is used to measure attitudes, opinions, and perceptions of an individual or group regarding social phenomena. The Likert scale scores range from 1 to 5, with a score of 1 being the lowest and 5 being the highest. Once validated by experts, the questionnaire is analyzed. The formula to calculate the average score per indicator is as follows:

$$Me = \frac{\sum X_i}{n} \quad (1)$$

where Me is the mean (average), $\sum X_i$ is the value of X from i to n, and n is the number of individuals. The percentage is obtained by

$$P = \frac{\sum x}{\sum x_i}\ x\ 100\ \% \quad (2)$$

where P is the percentage value, $\sum x$ is the total responses from respondents, and $\sum x_i$ is the total ideal value in the items.

Based on the obtained results, the percentage for validation criteria is calculated. This data analysis uses the Likert scale. The use of a Likert scale enabled the conversion of qualitative expert evaluations into quantifiable scores, facilitating structured comparison across criteria and supporting objective interpretation of instrument validity. The scale ranged from 1 (very inappropriate) to 5 (highly appropriate), allowing the identification of items requiring revision based on their relative rating levels. The validated questionnaire responses were analyzed quantitatively by calculating the mean score and percentage agreement for each evaluation criterion. These scores were then interpreted against predetermined feasibility thresholds to determine the level of appropriateness of each instrument component and to identify items requiring revision. The formula to calculate the average value per indicator is as follows:

$$Me = \sum Xn \quad (3)$$

Explanation:

To calculate the percentage of responses from respondents, the formula is as follows:

$$P = \sum x \sum x_i \times 100\% \quad (4)$$

where P is the percentage, $\sum x$ is the total responses from respondents, and $\sum x_i$ is the total ideal value in the items.

The validation data were converted into percentage agreement scores for each criterion to determine the feasibility level of the instrument. The interval classification and feasibility criteria were established using a criterion-referenced approach based on the proportion of the maximum possible score obtained from expert ratings. This approach enables standardized interpretation of validation outcomes by comparing the observed scores with the ideal score range, thereby allowing each percentage interval to represent a specific level of instrument appropriateness. The resulting percentage values were then interpreted according to

the predefined feasibility categories presented in Table 1.

Table 1. Interpretation Scale of Criteria

Interval (%)	Criteria
0 – 20	Not Good
21 – 40	Less Good
41 – 60	Fair
61 – 80	Good
81 – 100	Very Good

Construct Validity and Content Validity Test

This study evaluated instrument quality through construct validity, content validity, and reliability testing. Construct validity was examined to determine whether each test item appropriately represented the theoretical structure of conceptual understanding assessed through the CER framework. The evaluation focused on the alignment between item structure and the intended cognitive components, including clarity of conceptual claims, appropriateness of supporting evidence, and logical coherence of reasoning. Expert reviewers assessed each item using predetermined evaluation criteria, and their ratings were analyzed to identify items requiring revision in terms of structure, clarity, or conceptual representation.

Content validity was assessed to ensure that the instrument adequately represented the domain of work and energy concepts and corresponded to the intended learning indicators. Experts in physics education evaluated the relevance, accuracy, and representativeness of each item with respect to curriculum objectives and conceptual scope. The evaluation results were quantified and interpreted to determine the degree to which the instrument content reflected the targeted construct domain.

After establishing evidence of construct and content validity, the reliability of the instrument was examined to evaluate the internal consistency of students' responses. Reliability was estimated using Cronbach's Alpha coefficient, which measures the degree to which items within the instrument consistently assess the same underlying construct. The calculation of Cronbach's Alpha is presented in the following equation.

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right) \tag{5}$$

where

r_{11} : The reliability coefficient being calculated
n : The number of question items tested
$\sum \sigma_i^2$: The sum of the variance scores for each item
σ_t^2 : The total variance

Variance can be calculated, as referred to by (Arikunto, 2016).

$$\sigma_i^2 = \frac{\sum(Xi - X)^2}{N} \tag{6}$$

where

σ_i^2 : Item variance
N : The number of test participants
Σx^2 : The sum of the squared total scores

The difficulty level is used to determine whether a question is easy, moderate, or difficult. According to Arikunto (2016) the difficulty level of an item can be analyzed by examining the proportion of students who answer the item correctly, which indicates the relative ease or difficulty of the question.

$$\textit{The Difficulty Level (TK)} = \frac{Mean}{Maximum\ Score} \tag{7}$$

where:

Mean = (Total score for a specific question) / (Number of test participants)

As shown in Table 2, the difficulty level classification of the test items follows the criteria established by Arikunto (2016), which categorizes items into low, medium, and high difficulty levels.

Table 2. Difficulty Level Classification (Arikunto, 2012)

Limit (%)	Category
0.00 < TK ≤ 0.30	Difficult
0.31 < TK ≤ 0.70	Moderate
0.71 < TK ≤ 1.00	Easy

Table 2 displays the criteria for classifying item difficulty levels. These criteria were applied to the calculated difficulty index of each item to evaluate whether the test items were categorized as easy, moderate, or difficult.

Discrimination Power

The discrimination index was calculated to evaluate each item's effectiveness in differentiating students with higher and lower levels of conceptual understanding of work and energy. Students' total test scores were ranked and divided into upper and lower groups, and the proportion of correct responses in each group was compared to obtain the discrimination value for every item. Items with low or negative discrimination indices were flagged for revision or removal, while items with acceptable indices were retained for the final instrument. The discrimination power formula, is as follows:

DP = (Mean of the Upper Group - Mean of the Lower Group) / (Maximum Score)

As shown in Table 1.3, the discrimination power categories were used to evaluate the ability of each test item to distinguish between students with high and low levels of understanding

Table 3. Discrimination Power Categories

Limit	Category
0.00 < DP ≤ 0.19	Discarded
0.20 < DP ≤ 0.29	Needs Improvement
0.30 < DP ≤ 0.39	Acceptable, but needs improvement
0.40 < DP ≤ 1.00	Accepted

Functionality of Distractors

Distractor analysis was conducted to evaluate the quality of alternative responses in each item of the CER-integrated four-tier diagnostic test. For every item, the selection frequency of each distractor was calculated and examined across ability groups derived from total test scores. A distractor was considered functional when it attracted responses from at least 5% of participants and was predominantly selected by students in the lower-ability group. Distractors that were rarely selected or disproportionately chosen by higher-ability students were identified as non-functional and targeted for revision or replacement. This procedure was applied to ensure that all alternatives contributed effectively to diagnosing students' conceptual understanding of work and energy, following established item analysis guidelines (Arikunto, 2016).

Misconception Analysis for Students

The Confidence Discrimination Quotient (CDQ) is data used to analyze misconceptions based on the test results from the final field test. The data is analyzed to determine whether students can distinguish between what they understand and what they do not understand (Fariyani et al., 2015). According to Sugiyono (2017) , the data analysis for determining CDQ can be done using the following equation:

$$CDQ = (CFC - CFW) / S \quad (8)$$

Where:

CFC : Average confidence level of students who answered correctly

CFW : Average confidence level of students who answered incorrectly

S : Standard deviation of the confidence level

Interpretation of the Four-Tier Diagnostic Test Results

The interpretation of the Four-Tier Diagnostic Test is used to classify students into categories of understanding, misunderstanding, or misconceptions. The interpretation results are presented in a Table 4, which includes columns for answers, confidence levels of answers, reasoning, confidence levels of reasoning, and criteria. The guidelines for interpreting the Four-Tier Multiple Choice Diagnostic Test are adopted from Fariyani et al (2015) research. There are 16 possible criteria for students, as shown in Table 4.

Table 4. Four-Tier Diagnostic Test Interpretation

Answer	Confidence in Answer	Reasoning	Confidence in Reasoning	Criteria
Correct	High	Correct	High	Understand
Correct	Low	Correct	Low	
Correct	High	Correct	Low	
Correct	Low	Correct	High	
Correct	Low	Incorrect	Low	
Incorrect	Low	Correct	Low	Does Not Understand
Incorrect	Low	Incorrect	Low	
Correct	High	Incorrect	Low	Misconception
Incorrect	Low	Correct	High	
Correct	Low	Incorrect	High	
Correct	High	Incorrect	High	
Incorrect	High	Correct	Low	
Incorrect	High	Correct	High	
Incorrect	High	Incorrect	Low	
Incorrect	Low	Incorrect	High	

Scoring is given by assigning a score of 1 for correct answers or reasoning and a score of 0 for incorrect answers or reasoning. Confidence levels are considered high if rated 4, 5, or 6, and low if rated 1, 2, or 3 (Fariyani et al, 2015).

Student Responses Based on the Criteria of Understanding, Misunderstanding, or Misconception

Student responses are calculated as a percentage using the formula:

$$P = f/N \times 100\% \quad (9)$$

Where:

P = Percentage (%) of the group

f = Number of students in each group

N = Total number of individuals (total number of students in the study)

The percentage for each category (understanding, misunderstanding, or misconception) can be classified into several categories, as shown in Table 5.

Table 5. Categorization of Each Criterion (Suwarna, 2012)

Percentage (%)	Category
$0 < P \leq 30$	Low
$30 < P \leq 60$	Moderate
$60 < P \leq 100$	High

Results and Discussion

Following the data collection and analysis procedures described above, the findings are presented and summarized in Table 6.

Table 6. Results

Question	Understand (%)	Do Not Understand (%)	Misconception (%)
1	63	12	25
2	63	8	29
3	74	8	18
4	62	8	30
5	68	9	23

The present study examined students' conceptual understanding of work and energy using a four-tier diagnostic format integrated with CER. Across items, most twelfth-grade students demonstrated high levels of correct responses, yet misconceptions persisted consistently. This coexistence of correctness and conceptual error indicates that answer accuracy alone does not adequately represent the structure of students' knowledge. Instead, understanding appears partially constructed, with scientific and intuitive conceptions operating simultaneously within students' cognitive systems. Such a pattern aligns with the constructivist view that learning involves the gradual reorganization of prior knowledge rather than its immediate replacement (Posner et al., 1982).

The relatively high level of item difficulty observed in this study can be explained by both cognitive and epistemological characteristics of work–energy concepts. Unlike directly observable phenomena, energy is an abstract theoretical construct inferred through relationships among physical quantities. Students must coordinate force, displacement, and system conditions simultaneously, which increases intrinsic cognitive load. When cognitive demand exceeds processing capacity, learners tend to rely on intuitive heuristics rather than analytic reasoning. Because everyday experiences reinforce beliefs about effort, motion, and energy use, scientifically accurate explanations may appear less cognitively economical, thereby sustaining intuitive interpretations (Posner et al., 1982).

The multi-tier diagnostic structure employed in this study also contributes to increased item difficulty. Unlike conventional assessments that evaluate answer correctness alone, the instrument requires students to select an answer, justify reasoning, and express confidence. This layered structure raises the epistemic threshold for success by assessing not only knowledge but also reasoning coherence and metacognitive awareness. Therefore, high difficulty levels should not be interpreted solely as indicators of weak learning outcomes but also as evidence that the instrument effectively probes deeper conceptual structures.

The relatively low discrimination power of several items provides further insight into students' conceptual profiles. Discrimination indices depend on the extent to which items differentiate between high- and low-performing students. In contexts where misconceptions are widespread across achievement levels, statistical differentiation becomes limited. This phenomenon has been widely reported in physics education research. Studies conducted by David Hestenes demonstrated that students with varying performance levels frequently share the same alternative conceptions in foundational mechanics, indicating that conceptual misunderstanding is not restricted to low-achieving learners (Hestenes et al., 1992). The findings of the present study support this interpretation: low discrimination does not necessarily indicate weak item design but may instead reflect homogeneity in conceptual difficulty across learners.

Another factor contributing to low discrimination power is the separation between procedural competence and conceptual understanding. Students may correctly apply formulas while holding incorrect conceptual interpretations. The CER framework reveals such discrepancies by requiring justification of answers; however, when aggregated into item-level statistics, differences in reasoning quality may not produce strong separation between achievement groups. Thus, discrimination indices should be interpreted

cautiously within diagnostic assessment contexts where reasoning processes are emphasized.

The persistence of misconceptions across all items is consistent with the knowledge-in-pieces perspective proposed by Michelene T. H. Chi, which suggests that learners' conceptual systems consist of context-sensitive knowledge elements rather than unified theories. Misconceptions are therefore not random errors but stable patterns of reasoning activated in specific contexts (Chi, 2005). In work energy problems, students often interpret energy as a substance that is consumed or transferred, reflecting everyday experiential reasoning. Such ontological interpretations are cognitively intuitive and therefore resistant to change.

The present findings are also consistent with empirical investigations documenting persistent misunderstandings in work and energy. Research by Lillian C. McDermott demonstrated that many students who successfully solve quantitative problems still fail to interpret the physical meaning of work and energy relationships. Her work revealed a systematic gap between mathematical performance and conceptual understanding (McDermott, 1993). Likewise, reviews synthesized by Reinders Duit concluded that energy-related misconceptions are among the most stable conceptual difficulties in science learning due to their grounding in everyday language and experience (Duit, 2009). The patterns observed in this study corroborate these findings by showing that misconceptions remain present even when overall understanding appears high.

Beyond confirming prior literature, this study contributes by demonstrating how diagnostic instruments integrating reasoning and confidence can systematically detect conceptual instability that remains hidden in conventional testing. Traditional assessments often overestimate conceptual mastery by equating correctness with understanding. By contrast, the CER-integrated four-tier diagnostic test captures explanatory coherence, enabling a more accurate representation of students' conceptual structures. This methodological contribution aligns with contemporary perspectives on scientific literacy that emphasize reasoning as a core component of understanding (McNeill & Krajcik, 2012).

The observed conceptual patterns have direct implications for instructional practice, particularly within the context of SMAN 34 Jakarta. The distribution of responses indicates that instructional challenges lie primarily in restructuring incorrect knowledge rather than introducing entirely new concepts. Students are not devoid of information; rather, they interpret information through incompatible frameworks. Consequently, instruction should prioritize conceptual confrontation, explanatory dialogue, and representational integration.

Research in science education demonstrates that conceptual change is more likely when learners encounter cognitive conflict supported by evidence-based reasoning. Instructional strategies such as guided inquiry experiments, conceptual demonstrations, and structured argumentation tasks can facilitate this process. Within classroom practice at SMAN 34 Jakarta, teachers may incorporate CER-based discussions that require students to justify predictions, evaluate evidence, and revise explanations. Such practices align assessment with learning by making reasoning processes visible and supporting metacognitive reflection.

The findings also highlight the importance of representational diversity in teaching work and energy. Students often struggle to connect mathematical formulas with physical interpretations. Instruction that integrates graphical representations, real-world contexts, and experimental demonstrations can support conceptual coherence. When students translate among multiple representations, they are more likely to recognize inconsistencies in intuitive reasoning, thereby promoting conceptual restructuring (Ainsworth, 2006).

Another pedagogical implication concerns formative assessment. Diagnostic instruments such as the four-tier CER-based test should not function solely as evaluative tools but as sources of instructional feedback. By identifying specific misconception patterns, teachers can design targeted remediation activities. For example, if students interpret energy as "used up," instruction can explicitly address conservation principles through system-based reasoning tasks. Such targeted intervention is more effective than general review because it addresses the structure of misunderstanding.

Despite its contributions, the study has several limitations. The instrument assessed a limited number

of conceptual aspects within the work–energy domain, and the sample was drawn from a single educational context. Therefore, generalization to other topics or institutions should be undertaken cautiously. Additionally, the study employed a cross-sectional design that captures conceptual profiles at a single point in time. Longitudinal research is needed to examine how diagnostic-informed instruction influences conceptual change over time.

Future research should extend the application of CER-integrated diagnostic assessment to broader domains of physics and explore its impact on instructional decision-making. Investigating how teachers interpret diagnostic results and translate them into classroom practice would provide valuable insight into the relationship between assessment and pedagogy. Furthermore, combining quantitative diagnostic data with qualitative interview analysis could deepen understanding of students' reasoning processes.

In summary, high item difficulty reflects the cognitive demands of conceptual reasoning and the depth-oriented design of the diagnostic instrument. Low discrimination power indicates shared conceptual challenges across achievement levels rather than measurement weakness. Persistent misconceptions support theoretical perspectives emphasizing the stability of intuitive knowledge structures and the gradual nature of conceptual change. By integrating reasoning-based assessment with multi-tier diagnostics, this study provides a nuanced account of conceptual learning and offers practical directions for improving physics instruction, particularly within the learning environment of SMAN 34 Jakarta.

Conclusion

This study demonstrates that integrating Claim–Evidence–Reasoning (CER) into a four-tier diagnostic test provides a more informative assessment of students' conceptual understanding of work and energy than conventional formats that emphasize answer correctness alone. The instrument not only identifies whether students answer correctly but also reveals the structure and confidence of their reasoning, enabling the detection of stable misconceptions that might otherwise remain obscured. In this way, the FTDT + CER approach enriches physics assessment by linking conceptual accuracy with explanatory quality.

The persistence of misconceptions in work and energy indicates that students often rely on procedural knowledge without fully coordinating concepts, evidence, and principles. These findings highlight the need for instructional approaches that explicitly engage students in reasoning-based learning, such as conceptual conflict strategies, guided explanation, and structured argumentation. By making students' reasoning visible, the instrument offers actionable diagnostic information that can inform targeted remediation and support deeper conceptual change.

This study is subject to several limitations. The instrument was applied within a specific educational context and content domain, which may limit generalizability. In addition, the analysis relied on cross-sectional response data and did not examine changes in students' understanding over time or after instructional intervention.

Future research should investigate the use of CER-integrated diagnostic assessments across broader topics in physics, explore their impact on instructional decision-making, and examine how feedback based on diagnostic profiles supports conceptual change. Further refinement of item design is also recommended to enhance the instrument's discriminative capacity and optimize its role in classroom assessment practices..

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16(3), 183–198. https://doi.org/10.1016/j.learninstruc.2006.03.001

Arikunto, S. (2016). Research Procedures: A Practical Approach. Rineka Cipta.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. Journal of the Learning Sciences, 14(2), 161–199. https://doi.org/10.1207/s15327809jls1402_1

Duit, R. (2009). Students' and teachers' conceptions and science education. International Journal of Science Education, 31(3), 1–30. https://doi.org/10.1080/09500690803192650

Eliza, N., Tandililing, E., & Hidayatullah, M. M. S. (2022). Analysis of Students' Conceptual Understanding Abilities in Relation to Interest and Motivation in Learning Physics on the Topic of Work and Energy in Class X at SMA Negeri 6 Pontianak. Journal of Innovation in Research and Physics Learning, 3(1), 43. https://doi.org/10.26418/jippf.v3i1.49183

Fariyani, Q., Rusilowati, A., & Sugianto. (2015). Development of a Four-Tier Diagnostic Test to Uncover Misconceptions in Physics for High School Students in Class X. https://doi.org/10.1088/1742-6596/1567/2/022055

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. The Physics Teacher, 30(3), 141–158. https://doi.org/10.1119/1.2343497

Kalies, G. (2022). Back to the roots: the concepts of force and energy. Zeitschrift Für Physikalische Chemie, 236(4), 481–533. https://doi.org/10.1515/zpch-2021-3122

McDermott, L. C. (1993). Guest comment: How we teach and how students learn—A mismatch? American Journal of Physics, 61(4), 295–298. https://doi.org/10.1119/1.17228

McNeill, K. L., & Krajcik, J. (2012). Supporting grade 5–8 students in constructing explanations in science. Pearson.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception. Science Education, 66(2), 211–227. https://doi.org/10.1002/sce.3730660207

Setiawan, D., & Faoziyah, N. (2020). Development of a Five-Tier Diagnostic Test to Reveal Students' Concepts in Fluids. Physics Communication, 4(1), 6–13. http://journal.unnes.ac.id/nju/index.php/pc

Sheftyawan, W. B., Prihandono, T., & Lesmono, A. D. (2018). Identifying Students' Misconceptions Using a Four-Tier Diagnostic Test on Geometrical Optics Material. Journal of Physics Learning, 7(2). https://doi.org/10.19184/jpf.v7i2.7921

Sugiyono. (2017). Research and Development Methods. Alfabeta.

Sulistri, E., & Lisdawati, L. (2017). Using a Three-Tier Test to Identify the Number of Students Having Misconceptions on Newton's Laws of Motion Concepts. JIPF (Journal of Physics Education Science), 2(1). https://doi.org/10.26737/jipf.v2i1.195

Widayati, N. T., Wiyanto, W., & Subali, B. (2020). Analysis of the Quality of the Development of the Four-Tier Test to Determine the Profile of Metacognitive Ability of High School Students in the Ex-Karesidenan Pati Region. UPEJ Unnes Physics Education Journal, 9(2), 186–196. https://doi.org/10.1088/1742-6596/1460/1/012131

Wilantika, N., Khoiri, N., & Hidayat, S. (2018). Development of a Four-Tier Diagnostic Test Instrument to Uncover Misconceptions in the Material of the Excretory System at SMA Negeri 1 Mayong Jepara. Phenomenon Journal, 8(2). https://doi.org/10.1016/j.sbspro.2011.02.074