BIO EDUCATIO
(*The Journal of Science and Biology Education*)
http: https://jurnal.unma.ac.id/index.php/BE
p-ISSN : 2541-2280
e-ISSN : 2541-4097
DOI: http://doi.org/10.31949/be.v11i1.18083

Development of a Testlet-Based HOTS Assessment in Physics Using Polytomous Scoring Graded Response Model

Muhammad Nizar Ardhani[1]*, Elvin Yusliana Ekawati[2]
[1,2] Physics Education Study Program, Faculty of Teacher Training and Education,
Universita Sebelas Maret, Surakarta, Indonesia
Corresponding Author: nizarardhanimuhammad@student.uns.ac.id

Article Info
Article History
Received: 12-02-2026
Revised: 26-04-2026
Accepted: 30-04-2024

Keywords:

Higher-order thinking skills;
Testlet model;
Item response theory;
Graded response model;
Physics assessment

ABSTRACT
This study develops a higher-order thinking skills (HOTS) assessment instrument using a testlet model and examines its measurement characteristics in physics learning. A mixed-methods Research and Development (R&D) design was employed following Mardapi's framework, which includes planning, try-out, and measurement stages. Data were collected through expert validation, student responses, and field testing. Quantitative data were analyzed using the Graded Response Model (GRM), while qualitative analysis was used to evaluate content validity and item clarity. The results indicate that the instrument demonstrates acceptable measurement quality, with all items showing positive discrimination, appropriate difficulty levels, and adequate model fit. The testlet structure supports the assessment of sequential reasoning processes, allowing students' responses to be represented across different levels of understanding. These findings suggest that the instrument provides a valid and practical approach for assessing higher-order thinking skills in physics learning contexts.

INTRODUCTION

Assessment is an essential component of the learning process because it provides evidence for determining the extent to which learning objectives have been achieved. It includes the domains of attitudes, knowledge, and skills, while the knowledge domain is directly related to students' cognitive development. Cognitive competence generally covers remembering, understanding, applying, analyzing, evaluating, and creating. These levels are commonly classified as Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS). In contemporary education, HOTS has become a major priority because it supports critical thinking, creativity, and problem-solving competencies required in the twenty-first century (Widana, 2017; Yulianto et al., 2019). Learning environments that emphasize HOTS are more likely to prepare students for complex academic and social challenges. In Indonesia, the implementation of the Merdeka Curriculum also places strong emphasis on meaningful learning experiences and the development of HOTS-oriented competencies (Faradella et al., 2024).

The importance of HOTS is particularly evident in science and physics education. Physics learning requires students not only to understand formulas and concepts, but also to interpret phenomena, connect representations, and apply principles in unfamiliar situations. Effective

physics instruction, therefore, demands reasoning processes that integrate conceptual understanding with mathematical thinking (Kuo et al., 2013). Problem-based and inquiry-oriented learning approaches have also been shown to strengthen students' HOTS performance (Jailani et al., 2017; Wahyudi et al., 2022). However, in many classrooms, learning practices still emphasize procedural memorization rather than analytical reasoning. As a result, students often experience difficulty when confronted with contextual and multi-step physics problems (Eichenlaub & Redish, 2019).

Various international and national findings indicate that students' higher-order thinking skills remain relatively low. Large-scale assessments have consistently shown that student achievement in science and mathematics is still below expected international standards. These outcomes suggest that many students have not yet developed adequate abilities in reasoning, evaluating evidence, and solving non-routine problems (Thien et al., 2015). Similar patterns are reported in Indonesian classrooms, where assessment practices are often dominated by items that assess only recall and basic comprehension. Consequently, opportunities for students to practice HOTS through classroom assessment remain limited (Widana, 2017).

One major factor contributing to this condition is the limited availability of valid and practical assessment instruments specifically designed to measure HOTS. Many teachers still encounter difficulties in constructing items that assess analyzing, evaluating, and creating skills. Teacher-developed tests often focus on simple multiple-choice questions that are easier to prepare and score, yet less effective for measuring complex reasoning. This situation indicates the need for innovative assessment models that more accurately capture students' cognitive processes while remaining feasible for classroom implementation (Quaigrain & Arhin, 2017).

Several previous studies have attempted to develop HOTS-oriented instruments using research and development models such as ADDIE, Borg and Gall, 4D, and similar frameworks. These approaches have been widely applied to develop teaching materials, media, modules, and learning devices. Although useful, such models are generally broader in scope and not always specifically intended for psychometric instrument construction (Ediyanto et al., 2022). As a result, important aspects such as item functioning, scoring models, dimensionality, and response consistency may receive less attention. Therefore, a more focused framework is required to ensure that HOTS instruments are developed systematically and meet sound measurement standards (Maxnun et al., 2024).

One promising alternative is a testlet-based assessment model. A testlet consists of several interrelated items built around a common stimulus, scenario, or problem context. This structure enables students to solve problems progressively through a sequence of reasoning steps rather than responding to isolated items. Because responses reflect stages of thinking, testlets are considered highly relevant for measuring HOTS. In physics education, testlets are especially appropriate because many concepts require integrated reasoning across multiple sub-questions before arriving at a final solution. Recent developments in educational measurement also demonstrate an increasing use of testlet models to calibrate clustered items more accurately (Xiong et al., 2026).

To score testlet responses appropriately, Item Response Theory (IRT) models such as the Graded Response Model (GRM) can be employed. GRM is suitable for ordered response categories because it estimates student ability based on progressive achievement levels. Compared with conventional dichotomous scoring, GRM allows partial credit for intermediate reasoning stages and therefore provides richer information about students' cognitive

performance. This approach is particularly useful in HOTS assessment, where students may demonstrate partial understanding even when final answers are incomplete (Allen & Mattern, 2019; Marlina et al., 2025).

The suitability of the testlet model for HOTS assessment can be illustrated through the GRM scoring scheme presented in Table 1.

Table 1. Graded Response Model (GRM) Scoring Guidelines

No	Assessment Aspects	Score
1	The first item is answered incorrectly	0
2	The first item is answered correctly, but the second and third items are answered incorrectly or left unanswered	1
3	The first and second items are answered correctly, but the third item is answered incorrectly	2
4	All items are answered correctly	3
5	The first item is answered incorrectly, but the second or third item is answered correctly	0

Based on these considerations, this study aims to develop a higher-order thinking skills assessment instrument using the testlet model on motion dynamics material and to determine the characteristics of the resulting product. Motion dynamics was selected because it is conceptually demanding and requires students to integrate Newtonian principles, mathematical reasoning, and problem-solving strategies. Previous studies have shown that misconceptions and reasoning difficulties frequently occur in this topic (Rafika & Syuhendri, 2021; White, 1984). To ensure high measurement quality, the instrument development process employs a systematic framework emphasizing test construction, validation, reliability analysis, and item analysis (Mardapi, 2018; Ramadhan et al., 2019). Accordingly, the novelty of this study lies in integrating HOTS assessment, the testlet format, GRM scoring, motion dynamics content, and a specialized instrument-development framework to produce a valid, reliable, and practical physics assessment instrument.

METHOD

Research Design

This study employed a mixed-methods design, integrating quantitative and qualitative approaches to develop and test the instrument (Zhou, 2019; Deniz & Erdener, 2023). The quantitative component focused on analyzing item characteristics and estimating students' ability using Item Response Theory (IRT) with the Graded Response Model (GRM) (Mardapi, 2020; Saepuzaman & Istiyono, 2022; Zein & Akhtar, 2025). The qualitative component was used to assess content validity through expert judgment and to evaluate the instrument's readability and clarity based on students' responses during the pilot phase (Roebianto et al., 2023). This study is categorized as Research and Development (R&D) because it focuses on developing an assessment instrument (Umar et al., 2023; Mardapi, 2018).

Research Setting and Participants

This study was conducted in the Solo Raya region, Indonesia, involving senior high school students from several cities and regencies. The sample consisted of Grade XI students

from six schools selected to represent different geographical areas. A purposive sampling technique was used, with participants selected based on criteria relevant to the research objectives (Tajik et al., 2025). The main criterion was that students had completed instruction in motion dynamics, ensuring their prior knowledge aligned with the construct being measured.

Instruments

The instrument developed in this study was intended to measure higher-order thinking skills (HOTS) in motion dynamics using a testlet format (Merta Dhewa et al., 2017; Arafah et al., 2021). Each testlet was organized around a shared problem context and comprised a set of related items that required sequential responses, allowing students' reasoning processes to be observed more systematically (Ma et al., 2023). An initial problem was followed by several supporting questions designed to guide students through successive stages of thinking (Kocaoğlu & Şahin, 2024). The development process followed standard procedures in educational assessment, including test specification, item construction, validation, pilot testing, and analysis (Forde-Leaves et al., 2023). Content validity was examined through expert judgment, focusing on the material's relevance, the quality of item construction, and alignment with HOTS indicators (Roebianto et al., 2023). During the pilot phase, a readability questionnaire was used to identify potential issues related to clarity and comprehensibility (Tate et al., 2023). In addition, scoring was conducted using the Graded Response Model (GRM), which classifies responses into ordered categories for further analysis (Kurniawan et al., 2025).

Data Collection Procedure

The research procedure followed the stages of assessment instrument development proposed by Mardapi (2018), which consist of nine sequential steps. It began with a test specification to define the scope and indicators of the assessment, followed by item construction and review to examine their alignment with the intended constructs. The instrument was then implemented in a pilot phase to obtain empirical response data, which were analyzed to examine item characteristics. The results of this analysis informed revisions before the instrument was assembled for further use. Once assembled, the instrument was administered to participants, and the resulting data were scored and interpreted to estimate students' higher-order thinking skills. These steps were organized into three phases: the planning phase, focusing on test design and initial validation; the pilot and revision phase, involving empirical testing and refinement; and the measurement phase, covering test administration and interpretation of results, as presented in Figure 1.

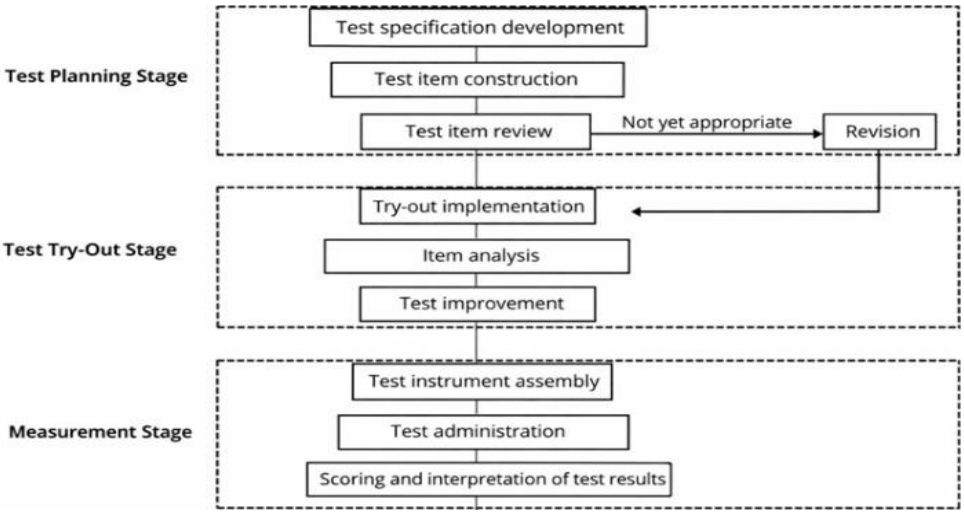

Figure 1. Mardapi's Test Instrument Development Procedure

Data Analysis

Quantitative Analysis

Quantitative data were analyzed using Item Response Theory (IRT) with the Graded Response Model (GRM) to examine the instrument's psychometric properties (Hori et al., 2022; Chen et al., 2025). The selection of GRM was aligned with the structure of the data, which consists of ordered response categories derived from the testlet format (Wallmark et al., 2024). The analysis began with evaluating the main assumptions underlying IRT, including unidimensionality to determine whether the instrument measures a dominant latent trait, local independence to assess whether responses to individual items are independent after controlling for students' ability, and parameter invariance to examine the stability of item and person parameters across different subsets of data (Yiğiter & Boduroğlu, 2024). Following the evaluation of these assumptions, item parameter estimation was conducted. Two primary parameters were estimated for each item, namely the discrimination parameter (α), which reflects the ability of an item to differentiate students across levels of ability, and the difficulty parameter (β), which represents the threshold levels of category difficulty (Gyamfi & Acquaye, 2023; Sweeney et al., 2022). These parameters were used to interpret item functioning and evaluate the instrument's measurement characteristics. The estimation process was conducted using the PARSCALE program.

Qualitative Analysis

Qualitative data were obtained through expert validation and student responses during the pilot phase and analyzed to assess the relevance of the content, the clarity of the language, and the instrument's usability. This analysis provided complementary evidence to the quantitative results by capturing how the instrument was interpreted and responded to by both experts and students. The analysis followed an interactive approach consisting of data condensation, data display, and interpretation (Miles et al., 2014). Data condensation involved selecting and organizing relevant feedback, particularly that related to alignment with higher-order thinking skill indicators and clarity of instructions. The organized data were then presented in a structured form to facilitate the identification of recurring patterns. The interpretation stage focused on identifying aspects requiring revision, including ambiguous wording, inconsistencies in item construction, and potential difficulties for students.

Ethical Considerations

This study adhered to established ethical standards for educational research involving human participants. Before implementation, permission was obtained from the relevant institutional authorities and course instructors. All participants were informed about the purpose, procedures, and voluntary nature of the study, and informed consent was secured before data collection. Students were assured that participation or withdrawal would not affect their academic evaluation or standing. To protect confidentiality, participant identities were anonymized using coded data, and all information was reported only in aggregate form. The intervention involved regular classroom learning activities and posed no physical or psychological risk. Both groups received equal access to instructional materials, lecturer support, learning duration, and assessment opportunities to ensure fairness. In addition, the researchers maintained academic integrity through accurate data reporting, transparent analysis, and proper acknowledgment of all cited sources.

RESULTS

The testlet-based higher-order thinking skills (HOTS) assessment instrument developed in this study was first reviewed by physics education lecturers and senior high school physics teachers. The validation results indicate that the instrument is appropriate in terms of content alignment, item construction, and language clarity. Feedback from the validators mainly addressed the wording of items and the contextual framing of problems to reduce ambiguity and improve relevance. The revisions carried out at this stage led to a clearer cognitive flow and closer alignment with HOTS indicators. Before proceeding to item analysis using Item Response Theory (IRT), the data were examined to ensure that key assumptions were met, including unidimensionality, local independence, and parameter invariance. Unidimensionality was assessed through factor analysis using SPSS. The results of the Kaiser–Meyer–Olkin (KMO) and Bartlett's Test (Table 2) show a KMO value of 0.932 and a significance level of 0.000, indicating that the data are suitable for further analysis.

Table 2. KMO and Bartlett's Test

Test Component	Indicator	Value
Kaiser-Meyer-Olkin Measure of Sampling Adequacy	KMO	0.932
Bartlett's Test of Sphericity	Approx. Chi-Square	2308.848
	Df	45
	Sig.	0.000

The results of the Kaiser–Meyer–Olkin (KMO) and Bartlett's Test (Table 2) indicate that the data are suitable for factor analysis, with a KMO value of 0.932, a chi-square value of 2308.848 (df = 45), and a significance level of 0.000, reflecting adequate sampling for further analysis. Unidimensionality was further examined through eigenvalue analysis. The eigenvalue plot (Figure 2) shows a single dominant component, as indicated by the steep decline after the first factor. This pattern suggests that a single underlying dimension underlies the data, indicating that the unidimensionality assumption is satisfied.

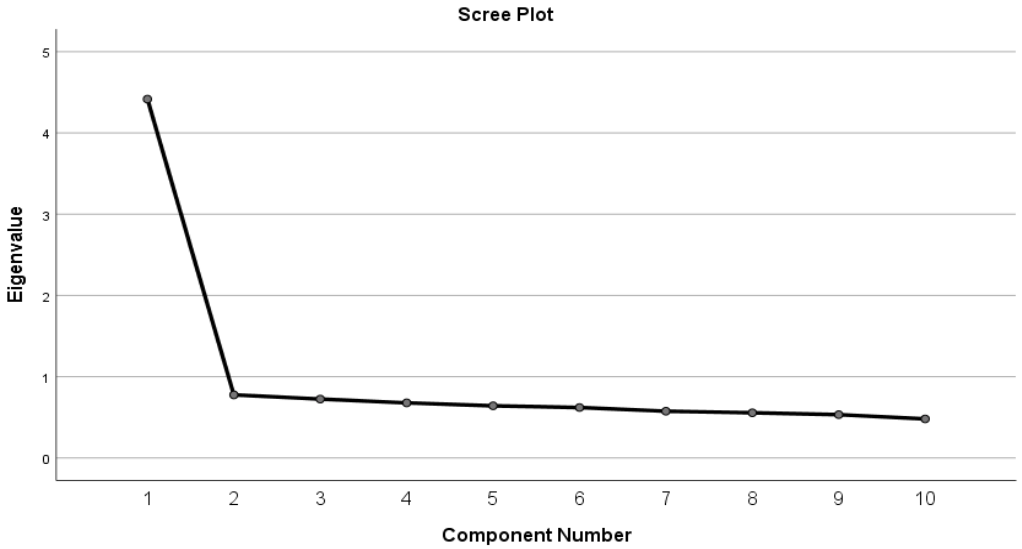

Figure 2. Eigenvalue Graph

Local independence was examined following the unidimensionality analysis. Given that the instrument demonstrated a dominant single factor, the assumption of local independence was considered satisfied, indicating that responses to each testlet were not influenced by responses to other testlets. Parameter invariance was then evaluated by comparing students' ability estimates derived from odd- and even-numbered items. This approach was used to examine the consistency of parameter estimates across different subsets of data.

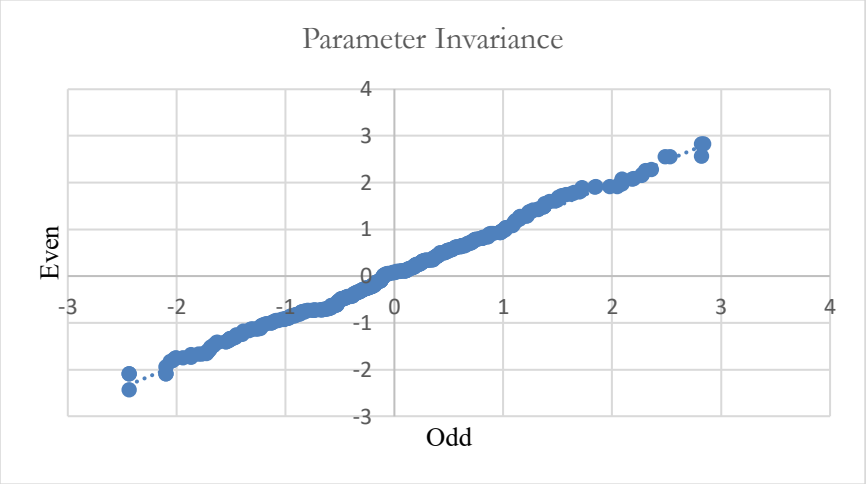

Figure 3. Parameter Invariance

Figure 3 shows that the distribution of points closely follows the line. $y = x$, indicating consistency between ability estimates derived from odd- and even-numbered items. This pattern suggests that the parameter invariance assumption is satisfied. The strength of this relationship is further supported by the correlation between the two sets of estimates, as presented in Table 3.

Table 3. Correlations Between Odd and Even Items

Variable	Statistic	ODD	EVEN
ODD	Pearson Correlation	1.000	0.937**
	Sig. (2-tailed)	—	0.000
	N	400	400
EVEN	Pearson Correlation	0.937**	1.000
	Sig. (2-tailed)	0.000	—
	N	400	400

The correlation between participants' ability estimates based on odd- and even-numbered items was 0.937, indicating a strong relationship. This result supports the fulfillment of the parameter invariance assumption. With all three assumptions satisfied, the instrument was considered appropriate for analysis using Item Response Theory (IRT). Item parameter estimation was subsequently conducted using the PARSCALE program.

Table 4. Test Item Parameter Analysis

Item	Block	Slope	S.E.	Location	S.E.	Guessing	S.E.
0001	1	0.784	0.036	-0.521	0.063	0.000	0.000
0002	1	0.852	0.037	-0.357	0.058	0.000	0.000
0003	1	0.948	0.042	-0.128	0.054	0.000	0.000
0004	1	0.729	0.032	0.109	0.065	0.000	0.000
0005	1	1.063	0.049	0.271	0.051	0.000	0.000
0006	1	0.854	0.039	0.531	0.059	0.000	0.000
0007	1	0.872	0.040	0.726	0.058	0.000	0.000
0008	1	0.868	0.043	0.907	0.060	0.000	0.000
0009	1	1.107	0.059	1.171	0.055	0.000	0.000
0010	1	0.940	0.050	1.403	0.063	0.000	0

The results of item parameter estimation indicate that all items exhibit positive discrimination values, suggesting that each item can differentiate students across varying ability levels. The discrimination parameters range from moderate to high, reflecting adequate sensitivity in distinguishing students' higher-order thinking skills. The difficulty parameters are distributed between −2 and 2, indicating that the items span difficulty levels from relatively easy to more challenging. This distribution allows the instrument to measure students' higher-order thinking skills across different ability levels.

Table 5. Model Fit Test

Block	Item	Chi-Square	D.F.	Prob.
BLOCK1	0001	13.13424	21	0.904
BLOCK1	0002	25.30241	21	0.234
BLOCK1	0003	25.48510	19	0.145
BLOCK1	0004	11.37736	21	0.955
BLOCK1	0005	36.47775	20	0.014
BLOCK1	0006	25.60135	21	0.222
BLOCK1	0007	29.61246	21	0.100
BLOCK1	0008	26.10849	22	0.247
BLOCK1	0009	22.01916	20	0.339
BLOCK1	0010	27.60629	21	0.151

Model fit was examined to assess the compatibility between the data and the Graded Response Model (GRM). Items are considered to fit the model when the probability value (p) is greater than 0.05. As shown in Table 5, most items have probability values exceeding 0.05, indicating a good fit to the model. However, one item (Item 0005) shows a probability value below the threshold ($p = 0.014$), suggesting potential misfit. Despite this, the overall results

indicate that most items conform to the GRM model, supporting the instrument's adequacy for measuring higher-order thinking skills.

DISCUSSION

The findings indicate that the developed instrument meets its intended purpose of measuring higher-order thinking skills (HOTS) in motion dynamics. Meaningful learning is closely associated with students' engagement in diverse thinking processes rather than routine or single-level tasks, as highlighted by Liljedahl (2021). This perspective is reflected in the item parameter estimates, in which all items show positive discrimination values, indicating their ability to differentiate students across varying ability levels. In addition, the distribution of difficulty parameters from -2 to 2 suggests that the instrument captures a broad spectrum of cognitive demand. Syamsi et al. (2025) further emphasize that effective assessment should capture variations in students' cognitive performance. Taken together, these results suggest that the instrument is sensitive to differences in students' reasoning and can capture variations in higher-order thinking, rather than focusing on single-level performance.

The results also show that the instrument satisfies the key assumptions of Item Response Theory (IRT), including unidimensionality, local independence, and parameter invariance. These assumptions provide a basis for interpreting the measurement results consistently. The presence of a dominant factor indicates that the instrument captures a single underlying latent construct, thereby supporting the interpretability of the estimated parameters (Reise et al., 2025). In addition, the close agreement between ability estimates derived from different item subsets suggests a stable estimation process, as fluctuations in ability estimates are often associated with measurement inconsistency (Hirose, 2023). A similar pattern is evident in the stability of item parameters, indicating invariance across conditions and supporting the credibility of the measurement outcomes (Feuerstahler, 2022). Overall, these findings point to internal coherence and stability in estimating students' abilities across different data conditions.

A clear distinction emerges when comparing this study with previous development research. Muna and Wardhana (2022) focus on the development of instructional media, particularly animation-based learning tools, while Purnamasari (2019) develops interactive media using Adobe Flash. Yusna et al. (2021) extend this line of work by developing learning devices to improve students' achievement, and Ryandhosi (2023) emphasizes module development by integrating research-based learning with local wisdom. Yuliaristiawan and Praherdhiono (2024) further highlight the use of digital platforms to support instructional models in classroom practice. Across these studies, the primary emphasis lies in producing effective and engaging learning tools, often accompanied by reported improvements in engagement and learning outcomes. However, limited attention is given to how students' higher-order cognitive abilities are systematically measured using validated instruments. In contrast, the present study focuses on developing an assessment instrument grounded in explicit psychometric properties, thereby addressing the measurement dimension of learning and complementing prior work.

The relationship between the obtained results and HOTS can be understood through both the instrument's structure and its theoretical basis. The use of testlet-based items encourages students to engage in sequential reasoning rather than responding to isolated questions. Such processes involve analysis, evaluation, and solution generation, reflecting the complex nature of higher-order thinking (Xiao et al., 2025). This interpretation is supported by Zhang et al. (2024), who highlight that HOTS emerges through interactive and context-based

reasoning processes. In addition, the variation in item difficulty indicates that the instrument presents different levels of cognitive demand. The items' ability to discriminate among students further suggests that differences in higher-level thinking are being captured. This is particularly relevant in science learning, where the development of higher-order thinking remains a persistent challenge (Akrami, 2022).

In addition to the instrument's structural design, the scoring approach plays an important role in representing students' cognitive performance. The application of the Graded Response Model (GRM) allows ordered response categories to be accounted for, enabling partial understanding to be reflected in the scoring process. This is particularly relevant in physics problem solving, where students may demonstrate intermediate reasoning even when final answers are incomplete. The use of GRM therefore provides a more detailed representation of student ability compared to dichotomous scoring. Empirical studies have shown that polytomous IRT models can improve the precision of parameter estimation, especially when responses reflect varying levels of understanding (Jiang et al., 2016). In addition, simulation-based evidence suggests that such models can produce stable estimates across different modeling conditions (Li, 2019). These findings support the use of structured scoring models to capture progressive cognitive performance.

The model fit analysis indicates that most items align with the GRM, although one item shows potential misfit and may require further review. This finding suggests that the instrument demonstrates acceptable measurement quality, while still allowing room for refinement. Revising misfitting items is a common step in assessment development, as empirical evidence is used to improve item functioning and alignment with the intended construct. The integration of HOTS constructs, testlet-based design, and GRM analysis reflects an effort to link theoretical considerations with empirical evidence. Each component contributes to representing students' cognitive performance in a more structured way. Future research may extend this approach to broader contexts, larger samples, and technology-based assessment environments that allow for more detailed response analysis.

CONCLUSION

This study developed a higher-order thinking skills (HOTS) assessment instrument using a testlet model in the context of motion dynamics. The results indicate that this approach captures students' reasoning processes more effectively. The integration of testlet design with Item Response Theory provides a consistent framework for representing higher-order cognitive performance. This study contributes by demonstrating how HOTS assessment can be operationalized through interconnected items and analyzed using a polytomous scoring model. Such an approach allows students' responses to be represented across different levels of understanding rather than as strictly correct or incorrect outcomes. In practice, the instrument offers a useful approach to evaluating higher-order thinking skills in physics learning contexts. However, this study is limited to a specific topic and sample, which may affect the generalizability of the findings. Future research may extend this approach to broader contexts, involve larger samples, and explore its application in digital or adaptive assessment environments.

ADDITIONAL INFORMATION

Section	Description
Funding	This research received no external funding.
Conflict of Interest	The authors declare no conflict of interest regarding the publication of this paper.
Author Contributions	All authors contributed significantly to this study. The first author was responsible for conceptualization, methodology, data analysis, and drafting the manuscript. The second author contributed to data collection, validation, and analysis. The third author reviewed, edited, and supervised the study. All authors approved the final manuscript.
Data Availability	The data supporting the findings of this study are available from the corresponding author upon reasonable request.
Ethical Approval	This study was conducted in accordance with ethical research standards. Informed consent was obtained from all participants, and confidentiality of the data was maintained.
AI Usage Statement	Artificial intelligence tools were used only for language refinement and grammatical editing. All research design, analysis, interpretation, and conclusions were carried out solely by the authors, who take full responsibility for the manuscript.

REFERENCES

Akrami, Z. (2022). The effectiveness of education with the STEM approach in the development of entrepreneurial thinking in chemistry students. *Chemistry Education Research and Practice, 23*(2), 475–485. https://doi.org/10.1039/D2RP00011C

Allen, J., & Mattern, K. (2019). Examination of indices of high school performance based on the graded response model. *Educational Measurement: Issues and Practice, 38*(2), 41–52. https://doi.org/10.1111/emip.12250

Arafah, K., Amin, B. D., Sari, S. S., & Hakim, A. (2021). The development of higher-order thinking skills (HOTS) instrument assessment in physics study. *Journal of Physics: Conference Series, 1899*(1), Article 012140. https://doi.org/10.1088/1742-6596/1899/1/012140

Chen, Y., Li, X., Liu, J., & Ying, Z. (2025). Item response theory—A statistical framework for educational and psychological measurement. *Statistical Science, 40*(2), 167–194. https://doi.org/10.1214/23-STS896

Deniz, Ü., & Erdener, M. A. (2023). Development and validation of the Trust in Higher Education Scale (THES): A mixed-methods approach. *Participatory Educational Research, 10*(3), 1–20. https://doi.org/10.17275/per.23.41.10.3

Ediyanto, E., Sunandar, A., Ramadhani, R. S., & Aqilah, T. S. (2022). Sustainable instrument development in educational research. *Discourse and Communication for Sustainable Education, 13*(1), 37–47. https://doi.org/10.2478/dcse-2022-0004

Eichenlaub, M., & Redish, E. F. (2019). Blending physical knowledge with mathematical form in physics problem solving. In G. Pospiech, M. Michelini, & B. S. Eylon (Eds.), *Mathematics in physics education* (pp. 101–120). Springer. https://doi.org/10.1007/978-3-030-04627-9_6

Faradella, N. E., Wiyaka, W., & Egar, N. (2024). Teachers' strategies and challenges in conducting HOTS-based activities in Merdeka curriculum era. *Indonesian Journal of Education and Pedagogy, 1*(2), 75–90. https://doi.org/10.61251/ijoep.v1i2.64

Feuerstahler, L. M. (2022). Metric stability in item response models. *Multivariate Behavioral Research, 57*(1), 94–111. https://doi.org/10.1080/00273171.2020.1868967

Forde-Leaves, N., Walton, J., & Tann, K. (2023). A framework for understanding assessment practice in higher education. *Assessment & Evaluation in Higher Education, 48*(8), 1076–1091. https://doi.org/10.1080/02602938.2023.2169659

Gyamfi, A., & Acquaye, R. (2023). Parameters and models of item response theory (IRT): A review of literature. *Acta Educationis Generalis, 13*(3), 68–78. https://doi.org/10.2478/atd-2023-0022

Hirose, H. (2023). Fluctuations of ability estimates in testing in item response theory. *International Journal of Learning Technologies and Learning Environments, 6*(1). https://doi.org/10.52731/ijltle.v6.i1.741

Hori, K., Fukuhara, H., & Yamada, T. (2022). Item response theory and its applications in educational measurement Part I: Item response theory and its implementation in R. *Wiley Interdisciplinary Reviews: Computational Statistics, 14*(2), e1531. https://doi.org/10.1002/wics.1531

Jailani, J., Sugiman, S., & Apino, E. (2017). Implementing the problem-based learning in order to improve the students' higher-order thinking skills and character. *Jurnal Riset Pendidikan Matematika, 4*(2), 247–259. https://doi.org/10.21831/jrpm.v4i2.17674

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology, 7*, 109. https://doi.org/10.3389/fpsyg.2016.00109

Kocaoğlu, S., & Şahin, M. G. (2024). Investigating the effect of testlets consisting of open-ended and multiple-choice items on reliability via generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 65–78. https://doi.org/10.21031/epod.1429423

Kuo, E., Hull, M. M., Gupta, A., & Elby, A. (2013). How students blend conceptual and formal mathematical reasoning in solving physics problems. *Science Education, 97*(1), 32–57. https://doi.org/10.1002/sce.21043

Kurniawan, N., Mujahidawati, M., & Falani, I. (2025). A modern approach to the accuracy of assessment of mathematical creative thinking ability through graded response models. *Journal of Educational Sciences, 9*(4), 2277–2288. https://doi.org/10.31258/jes.9.4.p.2277-2288

Li, T. (2019). A comparison between BMIRT and IRTPRO: A simulation study of a multidimensional item response model. *American Journal of Educational Research, 7*(11), 865–871. https://doi.org/10.12691/education-7-11-17

Liljedahl, P. (2021). Building thinking classrooms in mathematics, grades K–12: 14 teaching practices for enhancing learning. Corwin Press.

Ma, W., Wang, C., & Xiao, J. (2023). A testlet diagnostic classification model with attribute hierarchies. *Applied Psychological Measurement, 47*(3), 183–199. https://doi.org/10.1177/01466216231165315

Mardapi, D. (2018). Development of physics lab assessment instrument for senior high school level. *International Journal of Instruction, 11*(4), 17–28.

Mardapi, D. (2020). Assessing students' higher-order thinking skills using multidimensional item response theory. *Problems of Education in the 21st Century, 78*(2), 196–214. https://doi.org/10.33225/pec/20.78.196

Marlina, Hadi, S., & Rahim, A. (2025). The application of item response theory for developing higher-order thinking skills tests: A partial credit model study. *International Journal of Educational Reform*. Advance online publication. https://doi.org/10.1177/10567879251323884

Maxnun, L. L., Kristiani, K., & Sulistyaningrum, C. D. (2024). Development of HOTS-based cognitive assessment instruments: ADDIE model. *Journal of Education and Learning (EduLearn), 18*(2), 489–498. https://doi.org/10.11591/edulearn.v18i2.21079

Merta Dhewa, K., Rosidin, U., Abdurrahman, A., & Suyatna, A. (2017). The development of higher-order thinking skills (HOTS) instrument assessment in physics study. *IOSR Journal of Research & Method in Education, 7*(1), 26–32. https://doi.org/10.9790/7388-0701052632

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Muna, K. N., & Wardhana, S. (2022). Pengembangan media pembelajaran video animasi dengan model ADDIE pada pembelajaran bahasa Indonesia materi perkenalan diri dan keluarga untuk kelas 1 SD. *EduStream: Jurnal Pendidikan Dasar, 5*(2), 175–183. https://doi.org/10.26740/eds.v5n2.p175-183

Purnamasari, N. L. (2019). Metode ADDIE pada pengembangan media interaktif Adobe Flash pada mata pelajaran TIK. *Jurnal Pena SD, 5*, 23–31.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test. *Cogent Education, 4*(1), Article 1301013. https://doi.org/10.1080/2331186X.2017.1301013

Rafika, R., & Syuhendri, S. (2021). Students' misconceptions on rotational and rolling motions. *Journal of Physics: Conference Series, 1816*(1), Article 012016. https://doi.org/10.1088/1742-6596/1816/1/012016

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure higher-order thinking skills in physics. *European Journal of Educational Research, 8*(3), 743–751. https://doi.org/10.12973/eu-jer.8.3.743

Reise, S. P., Block, J. M., Mansolf, M., Haviland, M. G., Schalet, B. D., & Kimerling, R. (2025). Using projective IRT to evaluate the effects of multidimensionality on unidimensional IRT model parameters. *Multivariate Behavioral Research, 60*(2), 345–361. https://doi.org/10.1080/00273171.2024.2430630

Roebianto, A. D. I. Y., Savitri, S. I., Aulia, I. R. F., Suciyana, A. R., & Mubarokah, L. A. L. (2023). Content validity: Definition and procedure of content validation in psychological research. *Testing, Psychometrics, Methodology in Applied Psychology, 30*(1), 5–18. https://doi.org/10.4473/TPM30.1.1

Ryandhosi, A. (2023). Pengembangan modul berbasis model research-based learning terintegrasi kearifan lokal batik Incung. *Jurnal Vokasi Mekanika (VoMek), 5*(3), 294–300. https://doi.org/10.24036/vomek.v5i3.570

Saepuzaman, D., & Istiyono, E. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. *Pegem Journal of Education and Instruction, 12*(4), 269–279. https://doi.org/10.47750/pegegog.12.04.27

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). The relationship between IRT difficulty and discrimination. *Educational Measurement: Issues and Practice, 41*(4), 50–67. https://doi.org/10.1111/emip.12522

Syamsi, S., Junaidi, R., Hidayat, B., Syofyandi, R., & Julhadi, J. (2025). Evaluating the Quality of Objective and Essay Tests in Cognitive Assessment. El-Rusyd, 10(2), 140-146. https://doi.org/10.58485/elrusyd.v10i2.492

Tajik, O., Golzar, J., & Noor, S. (2025). Purposive sampling. *International Journal of Education & Language Studies*, 1–9. https://doi.org/10.22034/ijels.2025.490681.1029

Tate, R., Beauregard, F., Peter, C., & Marotta, L. (2023). Pilot testing as a strategy in questionnaire development. *Impacting Education, 8*(4), 20–25. https://doi.org/10.5195/ie.2023.333

Thien, L. M., Darmawan, I. G. N., & Ong, M. Y. (2015). Affective characteristics and mathematics performance. *Large-Scale Assessments in Education, 3*, Article 3. https://doi.org/10.1186/s40536-015-0013-z

Umar, U., Purwanto, M. B., & Al Firdaus, M. M. (2023). Research and development as an educational research framework. *Journal of English Language and Literature (JELL), 8*(1), 73–82. https://doi.org/10.37110/jell.v8i01.172

Wahyudi, W., Nurhayati, N., & Saputri, D. F. (2022). The effectiveness of problem solving-based optics module in improving HOTS. *Jurnal Penelitian Pendidikan IPA, 8*(4), 1992–2000. https://doi.org/10.29303/jppipa.v8i4.1860

Wallmark, J., Ramsay, J. O., Li, J., & Wiberg, M. (2024). Analyzing polytomous test data. *Journal of Educational and Behavioral Statistics, 49*(5), 753–779. https://doi.org/10.3102/10769986231207879

White, B. Y. (1984). Designing computer games to help physics students understand Newton's laws. *Cognition and Instruction, 1*(1), 69–108. https://doi.org/10.1207/s1532690xci0101_4

Widana, I. W. (2017). Higher-order thinking skills assessment. *JISAE: Journal of Indonesian Student Assessment and Evaluation, 3*(1), 32–44. https://doi.org/10.21009/jisae.v3i1.4859

Xiao, X., Li, Y., He, X., Fang, J., Yan, Z., & Xie, C. (2025). An assessment framework of higher-order thinking skills based on fine-tuned large language models. *Expert Systems with Applications, 272*, 126531. https://doi.org/10.1016/j.eswa.2025.126531

Xiong, J., Kuang, H., Tang, C., Liu, Q., Wang, B., Engelhard, G., Jr., Cohen, A. S., Xiong, X., & Sheng, R. (2026). A topic testlet model for calibrating constructed responses. *Journal of Educational Measurement, 63*, e70001. https://doi.org/10.1111/jedm.70001

Yiğiter, M. S., & Boduroğlu, E. (2024). Item response theory assumptions: A review. *International Journal of Educational Studies and Policy, 5*(2), 119–138. https://doi.org/10.5281/zenodo.14016086

Yulianto, T., Pramudya, I., & Slamet, I. (2019). Effects of 21st-century learning model on HOTS. *International Journal of Educational Research Review, 4*, 749–755. https://doi.org/10.24331/ijere.629084

Yuliaristiawan, E. D., & Praherdhiono, H. (2024). Penerapan model PLOMP berbantuan Padlet untuk pembelajaran IPA. *Journal of Educational Technology Studies and Applied Research, 1*(2), 1–7. https://doi.org/10.70125/jetsar.v1i2y2024a7

Yusna, L. M. A., Harjono, A., Ayub, S., & Wahyudi. (2021). Pengembangan perangkat pembelajaran dengan model reciprocal teaching untuk meningkatkan hasil belajar fisika

peserta didik. *KONSTAN: Jurnal Fisika dan Pendidikan Fisika, 6*(1), 72–78. http://jurnalkonstan.ac.id/index.php/jurnal

Zein, R. A., & Akhtar, H. (2025). Getting started with the graded response model. *International Journal of Psychology, 60*(1), e13265. https://doi.org/10.1002/ijop.13265

Zhang, M., Gou, J., Zhang, F., & Zhang, X. (2024). Assessment of higher-order thinking skills in conversation situation. *Procedia Computer Science, 242*, 845–852. https://doi.org/10.1016/j.procs.2024.08.210

Zhou, Y. (2019). A mixed methods model of scale development and validation. *Measurement: Interdisciplinary Research and Perspectives, 17*(1), 38–47. https://doi.org/10.1080/15366367.2018.1479088