Journal of Authentic Research
https://journal-center.litpam.com/index.php/jar/index

August 2026 Vol. 5, No. 3
e-ISSN: 2828-3724
pp. 6768-6780

Psychometric Analysis of a Complex Multiple-Choice Instrument for Measuring Critical Thinking Skills on Sound Waves Using Item Response Theory

Bayu Baihaqi[1*], Ika Mustika Sari[1], Duden Saepuzaman[1]

[1] Department of Physics Education, University of Education Indonesia, Indonesia.

*Corresponding Author e-mail: baihaqiibayu@upi.edu

Received: May 2026; Revised: July 2026; Published: August 2026

abstract

Critical thinking skills are an essential 21st-century competence, yet valid and reliable measurement instruments in physics remain limited in Indonesia. This descriptive quantitative study aims to analyze the psychometric characteristics of a complex multiple-choice instrument designed to measure students' critical thinking skills in sound waves. The study involved 187 eleventh-grade students from four senior high schools in West Java. The instrument consisted of 20 complex multiple-choice items developed based on Ennis's critical thinking framework (covering five aspects of critical thinking) and four subtopics of sound waves. Content validity was evaluated using Aiken's V index, while empirical validity was analyzed using the Rasch model. Item Response Theory (IRT) was applied by comparing the 1PL, 2PL, and 3PL models. The results showed that the instrument met the unidimensionality assumption with high content validity (Aiken's V = 0.939). Based on the Rasch model fit criteria, 19 out of 20 items (95%) were empirically valid. Among the IRT models, the 2PL model provided the best fit (AIC = 3637.95). The item difficulty parameters ranged from -3.878 to -0.153 logits, and the discrimination parameters ranged from 0.480 to 1.613. The test yielded a marginal reliability of 0.6946 and achieved maximum information at an ability level of θ = -1.55 logits, indicating that this instrument accurately measures students with low to moderate abilities. Overall, the instrument demonstrates satisfactory psychometric characteristics and stands as one of the first comprehensively developed tools to measure critical thinking in sound waves using IRT model comparison. Future studies are suggested to add items with higher difficulty levels to improve measurement accuracy for high-ability students.

Keywords: Critical Thinking Skills, Complex Multiple-Choice, Item Response Theory, Psychometric Analysis, Sound Waves.

How to Cite: Baihaqi, B., Sari, I. M., & Saepuzaman, D. (2026). Psychometric Analysis of a Complex Multiple-Choice Instrument for Measuring Critical Thinking Skills on Sound Waves Using Item Response Theory. *Journal of Authentic Research*, *5*(3), 6768–6780. https://doi.org/10.36312/jar.v5i3.6440

https://doi.org/10.36312/jar.v5i3.6440

INTRODUCTION

The rapid advancement of science and technology in the twenty-first century has transformed the competencies expected of students (Schwab, 2017). Beyond mastering scientific concepts, students are expected to develop higher-order thinking skills to analyze information, evaluate evidence, solve problems, and make reasoned decisions in various contexts (Johnson, 2010; Rachmadtullah, 2018). Among these competencies, critical thinking has become one of the most essential skills for science education because it supports students in understanding scientific phenomena and applying scientific knowledge to real-world situations (Mutiani & Faisal, 2019; Soh et al., 2010). In the context of physics learning, critical thinking enables students to evaluate experimental data, analyze relationships between variables, and apply scientific principles to everyday phenomena (Costa, 1985; Ennis, 2011).

The importance of critical thinking has been widely recognized in educational policies and international assessments (Kementrian Pendidikan dan Kebudayaan, 2017, 2022). However, the results of the Programme for International Student Assessment (PISA) 2022 indicate that Indonesian students continue to demonstrate relatively low performance in scientific literacy, suggesting that many students still experience difficulties in reasoning, evaluating evidence, and solving scientific problems. This low performance underscores the urgent need to develop assessment instruments capable of accurately measuring students' critical thinking skills (OECD, 2019; Sundari & Sarkity, 2021; Asniar, 2022).

Assessment plays an important role in determining whether students have achieved the intended learning outcomes (Arikunto, 2019; Nurgiyantoro, 2001). Nevertheless, many classroom assessments in physics still rely on conventional multiple-choice questions that primarily measure factual knowledge and procedural understanding (Fanani, 2018; Sugiarti et al., 2017). Such assessments provide limited opportunities for students to demonstrate critical thinking because they generally require selecting a single correct answer without evaluating multiple pieces of information or justifying reasoning processes (Anderson & Krathwohl, 2001; Wiggins & McTighe, 2005).

One alternative assessment format that has received increasing attention is the Complex Multiple-Choice (CMC) item (Albanese, 1993). Unlike conventional multiple-choice questions, CMC items require students to evaluate several statements simultaneously before determining the most appropriate combination of answers (Budiharti et al., 2023). This format reduces the probability of random guessing and encourages students to analyze evidence, compare alternatives, and make logical judgments (Haladyna & Rodriguez, 2013). Consequently, CMC items are considered more appropriate for assessing higher-order thinking skills, including critical thinking (Alosaimi et al., 2014).

Although CMC items offer considerable potential, developing an instrument alone is insufficient (Setyosari, 2016; Sukmadinata, 2012). The quality of the instrument must also be empirically demonstrated through rigorous psychometric evaluation (Retnawati, 2016). Classical Test Theory (CTT), which has traditionally been used in educational assessment, possesses several limitations because item statistics depend heavily on the characteristics of the sample used during testing (Sumaryanta, 2021). In contrast, Item Response Theory (IRT) provides sample-independent estimates of item parameters and allows researchers to evaluate item difficulty, discrimination, and measurement precision on a common latent ability scale (Hambleton & Swaminathan, 1985; Hambleton et al., 1991). Furthermore, IRT enables comparisons among different logistic models, including the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models, to determine the model that best represents the observed response data (Baker, 2001; Lord, 1980).

Within physics education, the topic of sound waves provides an appropriate context for assessing critical thinking because students are required to interpret physical phenomena, analyze relationships among variables such as frequency, wavelength, and wave speed, evaluate experimental data related to sound intensity and the Doppler effect, and apply scientific principles to everyday situations such as sound reflection and resonance (Anissa, 2020; Ennis, 2011). The topic encompasses concepts including the speed of sound, sound sources, sound intensity and intensity level, and the Doppler effect, all of which involve analytical reasoning beyond simple memorization (Demiral, 2018).

Despite the growing interest in developing critical thinking assessments in physics, a significant research gap remains. Studies specifically investigating the psychometric characteristics of complex multiple-choice instruments designed to measure critical thinking skills on the topic of sound waves using comprehensive IRT analysis—including systematic comparison of 1PL, 2PL, and 3PL models—are still limited (Handayani, 2022; Shanti et al., 2020). Several previous studies have developed or analyzed educational assessment instruments using Item Response Theory (Azizah & Wahyuningsih, 2020; Kurniawan, 2019; Pradani & Efendi, 2023). However, few have focused specifically on critical thinking in physics using complex multiple-choice formats with rigorous model comparison procedures.

To address this gap, the present study develops a CMC instrument for measuring students' critical thinking skills on sound waves and comprehensively evaluates its psychometric characteristics using Rasch analysis and Item Response Theory, including a comparison of the 1PL, 2PL, and 3PL logistic models. Specifically, this study examines the instrument construction, content validity, empirical validity, reliability, item difficulty, item discrimination, and the suitability of the 1PL, 2PL, and 3PL models to identify the model that best represents the response data. The findings are expected to contribute to the development of valid and reliable assessment instruments for measuring critical thinking skills in physics education, particularly through the systematic application of IRT model comparison procedures.

METHOD

This study employed a descriptive quantitative research design to analyze the psychometric characteristics of a critical thinking skills assessment instrument on the topic of sound waves (Sugiyono, 2017). The study focused on evaluating the instrument's construction, content validity, empirical validity, reliability, item difficulty, item discrimination, and overall measurement precision using Item Response Theory (IRT) (Nazir, 2014).

The participants consisted of 187 eleventh-grade science students from four senior high schools in West Java, Indonesia (from an initial 200 students, 13 students who obtained perfect scores were excluded because their ability estimates became infinite and could not be located on the logit scale, representing a standard procedure in Rasch analysis to ensure stable parameter estimation). The schools were selected using purposive sampling because the participants were required to have completed instruction on the sound wave topic before responding to the developed instrument (Sukmadinata, 2012). This criterion ensured that all students possessed the prerequisite conceptual knowledge needed to demonstrate their critical thinking skills during the assessment.

The research instrument was a 20-item Complex Multiple-Choice (CMC) test developed based on Ennis' critical thinking framework (Ennis, 2011). The instrument covered five aspects of critical thinking: (1) basic clarification, (2) basic support, (3) inference, (4) advanced clarification, and (5) strategies and tactics. The test items were distributed across four sound wave subtopics: speed of sound, sound sources, sound intensity and sound intensity level, and the Doppler effect (Anissa, 2020). The distribution of the items was designed to proportionally represent each critical thinking aspect while ensuring balanced coverage of the key concepts within the sound wave topic, thereby supporting comprehensive measurement of students' critical thinking skills.

Prior to field implementation, the instrument underwent content validation by three experts consisting of two physics education lecturers and one experienced physics teacher. Content validity was quantified using Aiken's V coefficient to evaluate the relevance of each item in terms of material, construction, and language (Aiken, 1985).

Empirical validity was evaluated using the Rasch model (1PL) through the eIRT software (Sumintono & Widhiarso, 2013). Item fit was determined based on Outfit Mean Square (Outfit MNSQ), Outfit Z-Standard (Outfit ZSTD), and Point Measure Correlation (PtMea Corr) (Fisher, 2007). In addition, the assumption of unidimensionality was examined using Principal Component Analysis (PCA) of residuals (Retnawati, 2016). Items satisfying the predetermined fit criteria were retained for subsequent analyses.

To identify the most appropriate IRT model, the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models were compared. Model selection was performed using the Akaike Information Criterion (AIC) and Likelihood Ratio (LR) tests (Hambleton et al., 1991). The AIC was used as the primary criterion to select the model with the best balance between model fit and parameter complexity, while the LR test was employed to determine whether adding parameters (such as discrimination in 2PL or guessing in 3PL) significantly improved the model fit (Veerkamp & Berger, 1999). After determining the best-fitting model, item difficulty (b) and discrimination (a) parameters were estimated. The guessing parameter (c) was also estimated for the 3PL model during model comparison.

The overall quality of the instrument was further evaluated through item reliability, person reliability, marginal reliability, Wright Map, Item Characteristic Curves (ICC), Item Information Functions (IIF), Test Information Function (TIF), and Test Characteristic Curve (TCC) (Baker, 2001; Retnawati, 2016). These analyses were conducted to determine the precision of the instrument across different levels of students' critical thinking ability and to identify items requiring revision or improvement.

The study acknowledges several limitations. First, the sample was limited to four senior high schools in West Java, Indonesia, which may restrict the generalizability of the findings to other regions with different educational contexts. Second, the instrument was developed specifically for the topic of sound waves, so the results may not be directly transferable to other physics topics without further validation. Third, the cross-sectional design provides a snapshot of instrument performance at one point in time, and longitudinal studies would be valuable to assess the stability of item parameters across different student cohorts.

RESULTS AND DISCUSSION

Instrument Construction

The developed instrument consisted of 20 complex multiple-choice (CMC) items designed to measure students' critical thinking skills on the topic of sound waves. The items were developed based on the five critical thinking aspects proposed by Ennis and distributed proportionally to represent each aspect. The distribution of the developed items is presented in Table 1.

Table 1. Distribution of the Developed Instrument Across Critical Thinking Aspects

Critical Thinking Aspect	Number of Items	Percentage (%)
Basic clarification	7	35.0
Basic support	4	20.0
Inference	5	25.0
Advanced clarification	2	10.0
Strategies and tactics	2	10.0
Total	**20**	**100.0**

Table 1 shows that the developed instrument places greater emphasis on the basic clarification aspect (35%), followed by inference (25%) and basic support (20%). The remaining items assess advanced clarification and strategies and tactics, each representing 10% of the total instrument. This proportional distribution reduces the likelihood that a single critical thinking aspect dominates the assessment, thereby providing a more comprehensive evaluation of students' critical thinking skills.

Content Validity

Before field testing, the developed instrument was evaluated by three experts consisting of two physics education lecturers and one experienced physics teacher. The experts assessed the relevance of each item in terms of material, construction, and language using Aiken's V coefficient. The summary of the content validity results is presented in Table 2.

Table 2. Summary of Content Validity Results

Aspect	Aiken's V	Category
Material	0.954	High
Construction	1.000	High
Language	0.833	High
Overall	0.939	High

Table 2 shows that the overall Aiken's V index of the developed instrument was 0.939, indicating high content validity. Among the three assessed aspects, the construction aspect achieved the highest agreement (1.000), followed by material (0.954), while language obtained the lowest value (0.833). Although the language aspect received the lowest score, it still met the high validity criterion, indicating that the instrument was appropriate for empirical testing after minor editorial revisions.

Empirical Validity and Unidimensionality

Following the content validation, the empirical validity of the developed instrument was evaluated using the Rasch model. Prior to the analysis, 13 students who obtained perfect scores were excluded because their ability estimates became infinite and could not be located on the logit scale. This exclusion represents a standard procedure in Rasch analysis to ensure stable parameter estimation. Consequently, the empirical validity analysis was conducted using responses from 187 students. Item validity was evaluated using three fit statistics: Outfit Mean Square (Outfit MnSq), Outfit Z-Standard (Outfit ZStd), and Point Measure Correlation (PtMeaCorr). An item was considered empirically valid if it satisfied at least two of the three fit criteria. The results of the empirical validity analysis are presented in Table 3.

Table 3. Summary of Item Fit Analysis

Category	Number of Items
Outfit (valid)	7
Fit 2 (valid)	12
Fit 1 (invalid)	1
Total Valid	19 (95%)

As shown in Table 3, seven items fulfilled all three fit criteria and were categorized as Outfit, whereas twelve items satisfied two of the three criteria and were categorized as Fit 2. Both categories were considered empirically valid. Only Item S7 met one criterion and was therefore classified as Fit 1 (invalid). Overall, 19 of the 20 developed items (95%) satisfied the Rasch item-fit criteria, indicating that the instrument demonstrated good empirical validity.

A closer examination of Item S7 revealed an Outfit MnSq value of 1.251, an Outfit ZStd value of 2.20, and a PtMeaCorr value of 0.258. The item failed to satisfy the Outfit ZStd and PtMeaCorr criteria, indicating an unexpected response pattern in which some high-ability students answered incorrectly while some low-ability students answered correctly. This finding suggests that Item S7 had limited ability to discriminate students according to their critical thinking ability and should therefore be revised or removed before future implementation.

The overall fit of the response data to the Rasch model was further supported by the mean Infit MnSq and Outfit MnSq values of 0.951 and 0.924, respectively. Both values fall within the acceptable range of 0.5–1.5, indicating that the overall response pattern was consistent with the expectations of the Rasch measurement model.

The unidimensionality assumption was examined using Principal Component Analysis (PCA) of standardized residuals. The results of the PCA analysis are presented in Table 4

Table 4. Summary of Unidimensionality Analysis

Component	Eigenvalue	Percentage (%)
Raw variance explained by measures	28.644	58.88
Raw unexplained variance (total)	20.000	41.12
Unexplained variance in 1st contrast	1.917	3.94
Unexplained variance in 2nd contrast	1.475	3.03
Unexplained variance in 3rd contrast	1.442	2.96

As shown in Table 4, the variance explained by the measures reached 58.88%, exceeding both the minimum criterion of 20% and the recommended value of 40%. Meanwhile, the unexplained variance in the first contrast was 3.94%, well below the maximum acceptable threshold of 15%. In addition, the first three contrasts all had eigenvalues below 2.0, indicating the absence of a meaningful secondary dimension. These findings confirm that the developed instrument satisfies the assumption of unidimensionality and measures a single latent construct, namely students' critical thinking skills.

IRT Model Selection

Following the empirical validity analysis, the most appropriate Item Response Theory (IRT) model was determined before estimating the item parameters. Three logistic models, namely the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models, were compared. Model selection was primarily based on the Akaike Information Criterion (AIC), followed by the Likelihood Ratio (LR) test and the Test Information Function (TIF). The summary of the model comparison is presented in Table 5.

Table 5. Summary of IRT Model Comparison

Criterion	1PL	2PL	3PL
AIC	3681.98	3637.95	3673.52
LR Test	–	Significant improvement over 1PL	No significant improvement over 2PL
Maximum Test Information	3.706	5.040	4.225
Decision		Selected	

As shown in Table 5, the 2PL model consistently demonstrated the best performance across the three model selection criteria. It produced the lowest AIC value, indicating the best balance between model fit and parameter complexity. The Likelihood Ratio test further showed that the 2PL model significantly improved the model fit compared with the 1PL model, whereas the addition of the guessing parameter in the 3PL model did not result in a statistically significant improvement. Furthermore, the 2PL model yielded the highest test information value, indicating greater measurement precision than the other two models. Therefore, the 2PL model was selected for estimating the item difficulty and discrimination parameters in the subsequent analyses.

Item Characteristics

Following the selection of the 2PL model, the item characteristics were evaluated based on the difficulty (b) and discrimination (a) parameters. The summary of the estimated item parameters is presented in Table 6.

Table 6. Summary of Item Parameters

Parameter	Minimum	Maximum	Interpretation
Difficulty (b)	−3.878	−0.153	Very easy to moderate
Discrimination (a)	0.480	1.613	Low to high

As shown in Table 6, the difficulty parameters ranged from −3.878 to −0.153 logits, indicating that all items were classified within the very easy to moderate categories (Hambleton et al., 1991). No items were categorized as difficult or very difficult. These findings suggest that the developed instrument was generally easy for the participating students. This result is consistent with the characteristics of the sample, where the average student ability exceeded the average item difficulty, as confirmed by the Wright Map analysis (Retnawati, 2016). The relatively low difficulty levels indicate that the instrument may not effectively challenge high-ability students, which has important implications for its use in differentiating students with higher critical thinking skills.

The discrimination parameters ranged from 0.480 to 1.613, indicating that all items possessed acceptable discrimination ability. Based on Baker's classification, the developed instrument consisted of items with low, moderate, and high discrimination levels, with most items falling within the moderate category. This indicates that the majority of the developed items were capable of distinguishing students with different levels of critical thinking ability.

Overall, the estimated difficulty and discrimination parameters indicate that the developed instrument possesses satisfactory psychometric characteristics. Nevertheless, the absence of difficult items suggests that additional items with higher difficulty levels should be developed to improve the measurement of students with higher critical thinking ability. This finding is consistent with the overall measurement precision observed in the subsequent reliability analysis.

Reliability and Measurement Precision

The quality of the developed instrument was further evaluated through Rasch reliability indices and the Test Information Function (TIF) based on the selected 2PL model. These analyses were performed to determine the consistency of the instrument and its measurement precision across different levels of students' critical thinking ability. The summary of the reliability analysis is presented in Table 7.

Table 7. Summary of Item Fit Analysis

Indicator	Value
Item reliability	0.936
Person reliability	0.546
Cronbach's alpha (KR-20)	0.683
Marginal reliability	0.6946
Maximum test information	5.040
Ability level at maximum information (θ)	−1.55

As shown in Table 7, the developed instrument demonstrated high item reliability (0.936), indicating that the relative difficulty of the test items was stable and could be replicated in different groups of students (Sumintono & Widhiarso, 2013). This high item reliability suggests that the item difficulty hierarchy is consistent and the instrument can be reliably used to rank items by difficulty across different student samples (Retnawati, 2016).

In contrast, the person reliability (0.546) and Cronbach's alpha (0.683) were considerably lower, suggesting limited consistency in distinguishing students based on their observed responses. However, this finding should not be interpreted as poor instrument quality. Instead, it reflects the mismatch between the difficulty level of the instrument and the ability distribution of the participating students. As indicated by the Wright Map analysis, the average student ability exceeded the average item difficulty, meaning that most items were relatively easy for the sample. Consequently, the instrument provided limited information for students with higher critical thinking ability, resulting in lower person reliability (Hambleton et al., 1991).

The Test Information Function further supports this interpretation. The instrument achieved its maximum information value of 5.040 at an ability level of $\theta = -1.55$, with a marginal reliability of 0.6946. These findings indicate that the developed instrument provides the highest measurement precision for students with low to moderate critical thinking ability, whereas its precision decreases for students with higher ability levels.

Comparison with Previous Studies

The findings of this study are consistent with previous research on critical thinking assessment in physics education. Sugiarti et al. (2017) developed a critical thinking assessment instrument for physics at the senior high school level and similarly found that Rasch analysis provided valuable information about item fit and person ability. However, their study primarily used CTT-based analysis without the systematic IRT model comparison employed in the present research.

Handayani (2022) analyzed the characteristics of physics test items using the Rasch model and reported comparable item reliability values, though their instrument focused on general physics content rather than specifically on critical thinking skills. Shanti et al. (2020) used the Rasch model with R software to analyze physics assessment items and emphasized the importance of evaluating item fit and unidimensionality, consistent with the approach taken in this study.

The unique contribution of the present study lies in the comprehensive comparison of 1PL, 2PL, and 3PL IRT models for a CMC instrument specifically designed for critical thinking in physics. While previous studies have used IRT for item analysis, few have systematically compared multiple IRT models to identify the best-fitting model for critical thinking assessments in physics (Kurniawan, 2019; Pradani & Efendi, 2023). The identification of the 2PL model as the best fit for this instrument extends the existing literature by demonstrating the importance of considering both difficulty and discrimination parameters when measuring higher-order thinking skills.

Furthermore, the moderate person reliability (0.546) and Cronbach's alpha (0.683) reported in this study are consistent with findings from Ariyanti & Bhakti (2020), who examined the reliability of multiple-choice tests in chemistry and similarly found that person reliability was influenced by the match between item difficulty and student ability. This underscores the need for targeted item development to improve measurement precision across the full ability range.

The high content validity achieved in this study (Aiken's V = 0.939) is comparable to that reported by Budiharti et al. (2023) in their development of complex multiple-choice instruments for creative thinking ability, suggesting that CMC formats with expert validation can achieve strong content validity when developed systematically.

CONCLUSIONS

This study successfully developed and validated a complex multiple-choice instrument for assessing critical thinking skills in physics, specifically on the topic of sound waves. The instrument effectively operationalized Ennis's five critical thinking aspects, demonstrating high content validity (Aiken's $V = 0.939$) and robust empirical validity, with 95% of items fitting the Rasch model and satisfying unidimensionality. Among the competing Item Response Theory models, the 2-parameter logistic model was identified as the most appropriate, confirming that both item difficulty and discrimination are essential parameters for accurately capturing the nuances of higher-order thinking. The high item reliability (0.936) further attests to the stability of the item difficulty hierarchy across different student populations.

However, the psychometric analysis revealed a significant limitation: the instrument primarily offers high measurement precision for students with low to moderate critical thinking abilities, as indicated by maximum test information at $\theta = -1.55$ logits. The uniformly negative difficulty parameters suggest that the test lacks challenging items, which explains the relatively low person reliability (0.546) and Cronbach's alpha (0.683) observed. Consequently, while the instrument is a valid and reliable tool for its intended lower-to-moderate range, it does not effectively discriminate among high-ability students. Overall, the instrument represents a valuable contribution to physics education assessment, providing a validated model for measuring critical thinking using complex multiple-choice formats, albeit with recognized targeting constraints that require further refinement.

RECOMMENDATIONS

Based on these findings, several recommendations are proposed for future research and practice. The misfitting item (S7) should undergo thorough revision by clarifying its wording and reviewing its keyed response to address the unexpected response patterns observed among high- and low-ability students. Crucially, to enhance the instrument's utility for high-performing students, additional items with higher difficulty levels must be developed to balance the test's targeting across the entire ability continuum and improve overall person reliability. Furthermore, the instrument should be empirically validated with larger and more diverse populations across different provinces in Indonesia to establish broader generalizability and confirm the stability of its psychometric properties. Finally, the application of this comprehensive IRT model comparison approach is encouraged for developing similar CMC instruments targeting critical thinking in other physics topics, such as electricity, mechanics, or thermodynamics, thereby expanding the repertoire of validated higher-order thinking assessments in physics education.

REFERENCES

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement, 45(1), 131–142. https://doi.org/10.1177/0013164485451012

Albanese, M. A. (1993). *Type K* and other complex multiple-choice items: An analysis of research and item properties. Educational Measurement: Issues and Practice, 12(1), 28–33. https://doi.org/10.1111/j.1745-3992.1993.tb00521.x

Alosaimi, K.H. & Reid, Norman & Rodrigues, S.. (2014). *Critical thinking*. Can it be measured?. Journal of Science Education. 15. 30-36.

Anderson, L. W. & Krathwohl, D. R., et al (Eds..) (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon. Boston, MA (Pearson Education Group)

Anissa, I. (2020). Modul Pembelajaran SMA Fisika: Gelombang Bunyi dan Cahaya. Probolinggo: SMAN 1 Probolinggo.

Arikunto, S. (2019). Prosedur Penelitian. Jakarta: Rineka Cipta.

Ariyanti, E., & Bhakti, Y. B. (2020b). Perbandingan Bentuk Tes Pilihan Ganda dan Teknik Penskoran Terhadap Reliabilitas Tes Mata Pelajaran Kimia. Titian Ilmu: Jurnal Ilmiah Multi Sciences, 12(2), 66–76. https://doi.org/10.30599/jti.v12i2.627

Asniar, A. (2022, August 15). Analisis Keterampilan Berpikir Kritis Dalam Pembelajaran Fisika Peserta Didik Di SMAN 11 Makassar. Jurnal Sains Dan Pendidikan Fisika. https://ojs.unm.ac.id/JSdPF/article/view/31622/16970

Azizah & Wahyuningsih, S. (2020). Penggunaan Model Rasch Untuk Analisis Instrumen Tes Pada Mata Kuliah Matematika Aktuaria. Jurnal Pendidikan Matematika, 3(1), 45-50. https://doi.org/10.30598/jupitekvol3iss1ppx45-50

Baker, F. B. (2001). The basics of *item response theory* (2nd ed.). College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Budiharti, R. D., Susilaningsih, E., Haryani, S., & Wardani, S. (2023). Pengembangan Instrumen Pilihan Ganda Kompleks Bermuatan Indikator Creative Thinking *Ability* Pada Materi Larutan Penyangga. Journal of Chemistry in Education, 12(2), 189-197. https://doi.org/10.15294/chemined.v12i2.69859

Costa, A. L. (1985). Developing of Minds (A Resource Book for Teaching Thinking). Washington DC: ASCD.

Demiral. (2018). Berpikir Kritis (Cara Jitu Menghadapi Tantangan Pembelajaran Abad 21). Jawa Timur: Gracias Logis Kreatif.

Ennis, R. H. (2011). The nature of *critical thinking*: An outline of *critical thinking dispositions* and abilities. University of Illinois, 2(4), 1-8.

Fanani, M. Z. (2018). Strategi Pengembangan Soal HOTS Pada Kurikulum 2013. Edudeena: Journal of Islamic Religious Education, 2(1), 57-76. https://doi.org/10.30762/ed.v2i1.582

Fisher, W. P. (2007). Rating scale instrument quality criteria. Rasch Measurement Transactions, 21(1), 1095.

Haladyna, T.M., & Rodriguez, M.C. (2013). Developing and Validating Test Items (1st ed.). Routledge. https://doi.org/10.4324/9780203850381

Hambleton, R.K., & Swaminathan, H. & Rogers, H.J. (1991). Fundamental of *item response theory*. Newbury Park, CA: Sage Publication Inc.
Hambleton, R.K., & Swaminathan, H. (1985). *Item Response Theory* Principles and Application. Boston, MA: Kluwer Inc.
Handayani, T. (2022). Analisis karakteristik butir soal fisika menggunakan model Rasch. Jurnal Pendidikan Fisika Indonesia, 18(1), 15–25.
Johnson, E. B. (2010). Contextual Teaching and Learning: Menjadikan Kegiatan Belajar Mengajar Mengasikkan dan Bermakna. Bandung: Kaifa Learning.
Kementrian Pendidikan dan Kebudayaan. (2017). Pendidikan Karakter Dorong Tumbuhnya Kompetensi Peserta Didik Abad 21. [Online]. https://www.kemdikbud.go.id/main/blog/2017/06/pendidikan-karakter-dorong-tumbuhnya-kompetensi-siswa-abad-21. [12 Desember 2023].
Kementrian Pendidikan dan Kebudayaan. (2022) Asesmen Komptensi Minimum. [Online]. https://pusmendik.kemdikbud.go.id/an/page/news_detail/asesmen-kompetensi-minimum. [20 Juni 2024].
Kurniawan, D.D. (2019). Analisis Butir Soal Ujian Akhir Semester Matematika Dengan Teori Respon Butir. Jurnal Riset dan Konseptual. 4(2), 215-224.
Lord, F. M. (1980). Applications of *item response theory* to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.
Mutiani, & Faisal, M. (2019). Urgency of the 21st Century Skills and Social Capital in Social Studies. The Innovation of Social Studies Journal, 1(1), 1–11.
Nazir, M. (2014). Metode Penelitian. Jakarta: Ghalia Indonesia.
Nurgiyantoro, B. (2001). Penilaian dalan Pengajaran Bahasa dan Sastra. Yoggyakarta: BPFE.
Pradani, R. A., & Efendi, A. (2023). Analisis Butir Soal Ujian Sekolah Menggunakan Program Iteman. Indonesian Language Education and Literature, 8(2), 276-289. https://doi.org/10.24235/ileal.v8i2.11002
Rachmadtullah, R. (2018). Pembelajaran Pendidikan IPS (Teori Konsep dan Aplikasi Bagi Guru dan Mahasiswa). Bekasi: CV Nurani.
Retnawati, H. (2016). Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian). Yogyakarta: Parama Publishing.
Schwab, K. (2017). The Fourth Industrial Revolution. Currency.
Setyosari, H. P. (2016). Metode penelitian pendidikan & pengembangan. Prenada Media. [Halaman 49-50].
Shanti, M. R. S., et al. (2020). Analisa Penilaian Soal Fisika Menggunakan Model Rasch Dengan Program R. Jurnal Sains dan Edukasi Sains, 3(2), 46-52. https://doi.org/10.24246/juses.v3i2p46-52
Soh, T., Arsad, N., & Osman, K. (2010). The relationship of 21st century skills on students' attitude and perception towards physics. Procedia Social and Behavioral Sciences, 7(C), 546–554.
Sugiarti, T., Kaniawati, I., & Aviyanti, L. 2017. Development of Assessment Instrument of *Critical Thinking* in Physics at Senior High School. Journal of Physics: Conf. Series 812 (2017) 012018.
Sugiyono. (2017). Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung: PT Alfabet.

Sukmadinata, N.S. (2012). Metode Penelitian Pendidikan. Bandung: PT. Remaja Rosdakarya.

Sumaryanta. (2021). Teori Tes Klasik & Teori Respon Butir: Konsep & Contoh Penerapannya. Cirebon: Confident.

Sumintono, B., & Widhiarso, W. (2013). Aplikasi Model Rasch untuk Penelitian Ilmu-Ilmu Sosial. Jakarta: Trim Komunikata Publishing House.

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. https://doi.org/10.24014/jnsi.v4i2.11445

Veerkamp, W. J. & Berger, M. P. (1999). Optimal item discrimination and maximum information for logistic IRT models. Applied Psychological Measurement, 23(1), 31-40.

Wiggins, G., & McTighe, J. (2005). Understanding by design (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development ASCD