JURNAL PENDIDIKAN MATEMATIKA DAN IPA

Vol. 17, No. 2 (2026) h. 311-322

http://jurnal.untan.ac.id/index.php/PMP

ANALYSIS OF THE VALIDITY OF MULTIPLE-CHOICE TEST ITEMS ON MASTERY OF TEMPERATURE AND HEAT CONCEPTS FOR SENIOR HIGH SCHOOL STUDENTS (SMA)

Jamiatul Ummah[1], Sutopo[2], Nandang Mufti[3], Sarah Lutfiah Ardilla[4]

[1234]Department of Physics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Malang, Jl. Cakrawala No. 5, Sumbersari, Lowokwaru District, Malang City, East Java, Indonesia 65145

Email Corresponding: jamiatul.ummah.2303218@students.um.ac.id

DOI: https://doi.org/10.26418/jpmipa.v17i2.95903

Abstract

This study aims to examine a test instrument to ensure its conformity with appropriate formats and procedures for measuring senior high school students' mastery of temperature and heat concepts. The instrument consists of 14 multiple-choice questions. This study uses a quantitative descriptive method to evaluate and produce a high-quality and valid test instrument. The instrument underwent revision through consultation with expert lecturers and was then empirically tested on 95 students at SMAN 1 Pamekasan as respondents. The test items were analyzed using SPSS version 30, covering four stages of empirical analysis: validity testing, reliability testing, item difficulty analysis, and item discrimination analysis, all conducted using classical test theory parameters. Based on the results of the point biserial analysis, questions 1, 9, and 11 are below standard. Reliability shows a Cronbach's alpha of 0.708, which means it is reliable. The difficulty level test results show that out of 14 questions, 3 questions are categorized as difficult and 1 question is very easy. In the discrimination test, 5 questions are categorized as poor, 3 questions are moderate, and 7 questions are good. Overall, 11 questions were deemed acceptable and passed the empirical test, while 3 questions (numbers 1, 9, and 11) were disqualified. The weakness of this instrument testing is that the results are highly dependent on the abilities of the participants. For example, question number 9, which was adapted from PhysPort and simplified in language, still had to be disqualified because it did not meet the empirical criteria.

Keywords**:** *Test Items; Multiple Choice; Concept Mastery; Senior High School*

Received **: 2025-07-04**
Revised **: 2025-09-02**
Accepted **: 2026-03-02**

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

INTRODUCTION

Numerous studies have examined students' conceptual understanding of temperature and heat. However, many findings in the field indicate persistent problems. For example, among six physics topics investigated in a study [1] namely kinematics, elasticity and Hooke's law, static fluids, optics, and temperature and hea-students' understanding was found to be lowest in the concepts of temperature and heat. Furthermore, students are often found to experience misconceptions regarding the concept of temperature and the effect of heat on thermal expansion related to changes in an object's dimensions [2]. This is caused by their failure to connect prior knowledge with new concepts and by possessing fragmented knowledge, such as merely memorizing facts without truly understanding the relationships among concepts [3], [4]. Students are also less skilled in analyzing physical quantities and in operating equations related to temperature and heat [4], [5]. In addition, limited understanding arises from students' inability to relate observable phenomena to the underlying scientific concepts [6]. These conditions have implications for students' mastery of concepts.

Students' mastery of concepts can be achieved when teaching strategies—such as methods, learning models, approaches, and evaluation materials for each meeting—are appropriately designed. Evaluation serves as a benchmark for determining the extent to which learning objectives have been successfully achieved [7]. According to the Merdeka Curriculum, evaluation materials in current learning practices consist of three components: formative assessment, summative assessment, and assessment instruments [8]. In practice, educators sometimes use self-developed instruments without knowing the quality of those instruments [9]. Although assessment standards have been established, in reality there are still a number of schools that have not successfully achieved these assessment objectives. One contributing factor is the unavailability of standardized assessment instruments [10], [11]. The use of assessment instruments that do not meet established standards will reduce the accuracy of results or the achievement of learning objectives. This condition makes it difficult for educators to distinguish between students who have achieved the required competencies and those who have not [10]. Consequently, when students' concept mastery or learning outcomes fail to meet competency standards, the problem may lie in the test instrument itself.

The type of measurement commonly used to determine students' mastery of concepts is multiple-choice testing, as employed in [3]. Multiple-choice items are test instruments that are widely used in formal assessments, for example in the administration of final examinations to measure students' learning outcomes [12]–[14]. However, the development of instruments to measure senior high school students' concept mastery of temperature and heat in the form of multiple-choice items is still limited, resulting in a small variety of test items for each indicator. Although there are validated and high-quality items available for the senior high school level, such as those provided by PhysPort, they are limited to thermal

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

and thermodynamics themes and do not specifically address instruments that examine the conceptual effects of heat on thermal expansion and Black's principle. Other studies have also been conducted, such as [15], but the development of instruments in those studies focuses more on classical thermodynamics and statistical mechanics at higher education levels.

The implementation of final assessment in learning aims to evaluate the learning process and to identify deficiencies that need to be improved in subsequent instruction [7]. Therefore, the use of well-designed test instruments is essential for measuring students' abilities. This process is carried out to determine the characteristics of each test item and to assess whether it is suitable for use through systematic analysis [7], [16]. In addition, item analysis enables the evaluation of the extent to which the quality of test items measures the various aspects of the assessed competencies [7]. Consequently, item analysis constitutes an important step in producing assessment instruments that are valid, accurate, and capable of appropriately reflecting students' abilities.

Based on previous studies, this research aims to examine the validity, reliability, difficulty level, and discrimination index of the developed multiple-choice instrument on temperature and heat. A valid instrument is able to measure the competencies that are intended to be assessed in learning [14], [17], [18]. The reliability of a test item is reflected in the consistency of students' test results across different occasions [14], [19]. The level of difficulty should be balanced, meaning that items should be neither too easy nor too difficult [20], [21]. The use of appropriate, clear, and simple language that does not lead to multiple interpretations or confusion also greatly influences item quality [22], [23]. Therefore, it is important to conduct testing of the developed instrument to ensure that it meets the criteria of validity and reliability, as well as to determine the level of difficulty and the discrimination power of each item. Instruments that have been proven to be valid and reliable will yield high-quality test items, thereby enabling accurate measurement of students' concept mastery. It is important to note that high-quality items do not function solely as tools for evaluating learning outcomes, but can also help strengthen students' conceptual understanding. Through items with proportional levels of difficulty and good discrimination power, students are encouraged to think more deeply, connect the concepts they have learned, and identify errors or misconceptions they may hold.

METHOD

This study is quantitative in nature and aims to provide an evaluation and produce a high-quality concept mastery test instrument suitable for general use. The test consists of 14 multiple-choice items, which were administered to students at SMAN 1 Pamekasan who had completed the temperature and heat material. The study involved 95 respondents from three eleventh-grade classes, with each class comprising approximately 30 to 35 students. A sample size of around 30–100 respondents is considered sufficiently accurate for instrument testing [24].

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

Prior to administration, the test items underwent validation by expert physics lecturers through consultation and several stages of revision. The test items were designed to measure students' concept mastery in the topic of temperature and heat. A description of each test item is provided in Table 1.

Table 1. Description of Test Items on Concept Mastery of Temperature and Heat

Topic	Item	Question Content	Cognitive Level
Temperature and Expansion	1	Understanding the effect of low atmospheric pressure on the boiling point of water and its impact on making tea in mountainous areas	C2
	2	Distinguishing between the expansion of a solid and a hollow sphere when subjected to the same temperature change	C2
	3	Explaining the process of opening a metal bottle cap through heating	C2
Heat and Its Relationship with Temperature and State Changes	4	Explaining the relationship between the amount of heat energy absorbed by an object and the specific heat capacity of that object	C3
	5	Determining which of two objects with different properties has a higher specific heat capacity based on the comparison of their temperature rise rates when heated simultaneously	C4
	6	Explaining the effect of pressure on the boiling point of water based on the process of cooking with a pressure cooker	C2
Black's Principle	7	Understanding the main idea of Black's principle (two objects with different temperatures mixed or in thermal contact)	C2
	8	Applying Black's principle to determine the temperature of two substances in thermal contact	C3
	9	Analyzing the conservation of heat energy to determine the final temperature of a mixture of two substances with different masses and initial temperatures in an isolated system	C3
Heat Transfer	10	Explaining the structure of a thermos bottle to prevent rapid temperature loss	C2
	11	Understanding the difference in the sensation of cold on objects made of different materials based on the concepts of thermal conductivity and heat transfer rate	C2
	12	Explaining the rationale for placing an air conditioner at the top of a room based on the principle of heat transfer by convection	C2

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

Topic	Item	Question Content	Cognitive Level
	13	Explaining the cause of temperature increase on asphalt surfaces exposed to sunlight based on the process of heat transfer by radiation	C2
	14	Explaining the function of thick curtains in inhibiting heat transfer through conduction and radiation to maintain room temperature stability	C2

The empirical testing of the items was analyzed using classical test theory parameters with the assistance of SPSS version 30, including the point-biserial correlation coefficient (r_{pbi}), the discrimination index (D), and the difficulty index (p) for each item. Validity testing was conducted using the point-biserial coefficient, examined through Pearson correlation and significance values to determine the level of validity. Given the sample size of 95 students, the critical r-value used was 0.202 at a 5% significance level (0.05) [25]. Reliability testing was assessed using Cronbach's alpha to determine the accuracy and consistency of the test instrument in measuring the dependent variable, ensuring that it can be used repeatedly. Item discrimination was evaluated using the Corrected Item–Total Correlation for each item. Item difficulty was assessed by dividing the mean score by the maximum score. The final decision on whether an item is suitable for distribution as a measure of students' concept mastery depends on whether it meets the parameter criteria, as presented in Table 2.

Table 2. Test Eligibility Parameters

Classical Test Parameters	Eligibility Criteria
Point-Biserial Correlation	≥ 0,2
Reliability	≥ 0,7
Discrimination Index	≥ 0,3
Difficulty Level	0,3 - 0,9

[21]

2. A ring and a solid ball are made of the same metal. At first, the solid ball barely fits into the ring. In order for the ball to fit into the ring, which alternative is the most effective?

Image source: https://youtu.be/GgrnhHMqZDY?si=oQnZj_58ZYzU4OQJ

A. The ring is heated, but the solid ball is left alone.
B. The solid ball is cooled, the ring is left alone.
C. The ring is heated and at the same time the solid ball is cooled.*
D. The ring is cooled and at the same time the solid ball is heated.

Figure 1. Example of a Concept Mastery Test Item on Temperature and Heat

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

RESULT AND DISCUSSION

The empirical testing of the concept mastery test instrument involved 14 multiple-choice items, which were administered to 95 eleventh-grade senior high school students who had completed the temperature and heat material. Before presenting the analysis results, the researcher provides examples of several tested items in Figure 1, and access to the complete set of instruments is available at the following link: https://drive.google.com/file/d/1Y0bYBTD1iWDed4N16eupyELlC3TtaJBj/view?usp=sharing. The results of the classical test parameter analysis are presented in Table 3.

Table 3. Results of Classical Test Parameter Analysis for the Concept Mastery Test on Temperature and Heat

Item	D	p	r_{pbi}
1	0.088	0.94	0.19
2	0.2	0.41	0.277
3	0.409	0.49	0.491
4	0.387	0.53	0.471
5	0.563	0.41	0.629
6	0.2	0.31	0.270
7	0.344	0.47	0.431
8	0.473	0.52	0.550
9	0.009	0.12	0.174
10	0.521	0.25	0.583
11	0.134	0.15	0.169
12	0.413	0.67	0.491
13	0.452	0.50	0.530
14	0.427	0.82	0.489

Point-Biserial Coefficient Test

Based on Table 3, r_{pbi} indicates the validity of each test item. The analysis results are presented in Table 4.

Table 4. Classification of r_{pbi} for Test Items

Number of items	Item numbers	Category
11	2, 3, 4, 5, 6, 7, 8, 10, 12, 13, & 14	Valid
3	1,9, &11	Not valid

Tables 3 and 4 show that 11 items passed and were classified as valid, while 3 items—numbers 1, 9, and 11—were found to be invalid. These three items cannot be used to measure students' concept mastery because their Pearson correlation values fall into the low category, meaning the calculated r-values are below the critical r-value, indicating that the items are of low quality and therefore unsuitable for measuring the dependent variable in this study. According to [21], [26], an instrument with low validity interpretation but a significance value below 0.05 is not

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

suitable for use, as the minimum acceptable validity interpretation should be at least "adequate" for the instrument to be usable. The invalid items—1, 9, and 11—confused students during completion [7]. This may be due to confusing item formats, such as improper wording, the presence of two correct answer options, or possibly students' low ability, and in some cases, the items may have been too easy or too difficult. Thus, multiple factors can contribute to an item's invalidity. Based on these possible factors, the researcher predicts that all of them could have occurred, affecting the items' validity, which is closely related to the difficulty level of the test items. Meanwhile, the 11 valid items were mostly classified as having "adequate" quality, with one item rated as high quality. This is because the item scores for each question were slightly lower than the total test score, resulting in the majority of valid items being rated as "adequate." As noted by [7], [27], there is a relationship between the item score and total score regarding the height or significance of the Pearson correlation and the validity of the item. If each item score aligns closely with the total score, the resulting validity is higher.

Reliability Test

The second stage, after conducting validity screening, is to ensure the reliability of the test instrument. The purpose is to confirm that the instrument produces consistent results when used, allowing for repeated measurements or use in the future and/or with different subjects [28]–[31].

Reliability Statistics

Cronbach's Alpha	N of Items
.597	14

Figure 2. Reliability Test Results for 14 Items

Reliability Statistics

Cronbach's Alpha	N of Items
.708	11

Figure 3. Reliability Test Results for 11 Items

The reliability test was analyzed twice to examine the difference in reliability before and after item deletion, following the prior validity testing. Figure 2 shows the results of the reliability test, indicating that the 14-item version of the test was categorized as unreliable, with a Cronbach's Alpha of 0.597. In contrast, the 11 valid items, after reliability testing, were classified as moderate and acceptable, with a Cronbach's Alpha of 0.708. This index value of 0.708 meets the desired threshold of ≥0.7, indicating that the 11 items generally measure the same concept—temperature and heat. This category is referred to as moderate reliability because it lies on the borderline between reliable and unreliable. The reliability index for the 11 items, classified as moderate, indicates a medium but adequate correlation among test items and consistency in measuring the concepts of temperature and heat [32], [33]. This demonstrates that the 11 items can be used to differentiate students' levels of mastery in the concepts of temperature and heat.

Item Difficulty Test

Based on Table 3, the analysis of the difficulty level for the 14 test items

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

shows that most items fall into the moderate category, meaning their difficulty is neither too high nor too low. The classification of item difficulty is presented in Table 4.

Table 4. Difficulty Level of Each Test Item

Number of items by item number	Category
2, 3, 4, 5, 6, 7, 8, 12, 13, & 14	Moderate
9, 10, &11	Difficult
1	Very easy

The analysis results show that three items—numbers 9, 10, and 11—fall into the difficult category, while one item is classified as very easy. These findings align with student feedback in the field, indicating that most items were perceived as having moderate to high difficulty. This is further supported by the students' responses, where many did not provide an answer or gave two answers for the same item, suggesting confusion in selecting the correct option in the multiple-choice questions.

Nevertheless, item number 10 was retained because it is part of the 11 items with acceptable reliability. Although its difficulty index is still 0.2, it is conventionally considered adequate and acceptable [33]. Moreover, retaining test items with higher difficulty levels has been applied in the development of KCT instruments [34]. As long as the items are neither extremely difficult nor extremely easy, they can still be used to provide variation in item difficulty [21]. Therefore, the inclusion of difficult items in an instrument can be maintained as long as it aligns with the objectives of the instrument's development [35].

Item Discrimination Test

Based on the data presented in Table 3, it can be concluded that the discrimination levels of the test items, in measuring concept mastery, range from 0.009 to 0.563. Analysis of the 14 items shows that they fall into three categories: poor, moderate, and good. Among all items, five were classified as poor (items 1, 2, 6, 9, and 11), two items (4 and 7) were classified as moderate, and seven items were classified as good.

These findings are in line with [7], who stated that items with very low discrimination should be removed or not reused. This is further supported by [7], [30], who noted that the higher the discrimination value, the greater the proportion of students able to answer the item correctly; conversely, a low discrimination value indicates that the item is more difficult and less effective in distinguishing students' abilities.

Based on this analysis, three items categorized as poor among the 14 test items were removed. Meanwhile, items 2 and 6 were retained because they remain on the borderline between poor and moderate categories. Another reason for retaining these items is that the instrument is intended to provide feedback to students and support improvements in the learning process, consistent with the findings of [35]. Therefore, it can be concluded that the test instrument suitable for use and meeting the discrimination criteria consists of 11 items, including those in the moderate category as well as items 2 and 6.

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

This study is valuable for producing higher-quality assessment instruments on the topic of temperature and heat at the senior high school level. The findings can serve as a reference for teachers and other researchers in designing test items that align with students' characteristics, ensuring that the instruments used are not only valid and reliable but also accurately reflect concept mastery. In the future, similar research can be extended to the same topics at different educational levels, providing comprehensive instruments tailored to specific learning needs.

However, a comprehensive analysis revealed a notable weakness in the testing of the instrument. According to classical test theory, when examining item difficulty and discrimination characteristics, a significant limitation was observed in this study. This weakness is related to the ability of the test takers, which aligns with the findings of [21]. Additional evidence of this limitation is seen in item number 9, originally taken from PhysPort and re-tested with language modifications to facilitate student comprehension. In practice, however, item 9 was found to be invalid, with its discrimination and difficulty indices rated as "very poor." This indicates that the issue was not with the quality or usability of the instrument itself, but rather depended on the students' abilities and skills.

CONCLUSION

The analysis of the 14 test items used to measure concept mastery, administered to 95 eleventh-grade students at SMAN 1 Pamekasan, concluded that 11 items were valid and suitable for use. The reliability analysis also showed a Cronbach's Alpha of 0.708, indicating that these 11 items are both valid and reliable. Among the 14 items, students experienced difficulty answering items 9 and 11, which were classified as very difficult based on the difficulty test. The item discrimination analysis categorized the 14 items into three groups: 5 items were poor, 3 items were moderate, and 7 items were good. Based on this analysis, only 11 items met the empirical testing criteria according to classical test parameters.

The implication of this study is that the 11 items can be used to measure students' concept mastery in future research and serve as assessment tools for evaluating learning. Additionally, the instrument can be further developed into other types of assessment instruments.

REFERENCES

[1] H. Hariyanto, "Kesulitan pemecahan masalah fisika pada siswa man 3 jember," *AS-SUNNIYYAH*, vol. 1, no. 02, pp. 58–67, 2021.

[2] E. W. N. Sofianto, R. K. Irawati, H. A. Akmalia, and R. R. Apriani, "The analysis of heat-temperature misconception and integration with Al Quran," in *Journal of Physics: Conference Series*, IOP Publishing, 2020, p. 12022.

[3] A. K. Hara, K. A. Astiti, and V. Lantik, "Analisis Penguasaan Konsep Fisika pada Materi Suhu dan Kalor Pasca Pembelajaran Online di Kelas XI SMA Negeri 12 Kota Kupang," *J. Ilmu Pendidik. STKIP Kusuma Negara*, vol. 14, no. 2, pp. 118–126, 2023, doi: 10.37640/jip.v14i2.1548.

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

[4] A. N. Laili, Sutopo, and M. Diantoro, "Ragam Kesulitan Siswa SMA dalam Menguasai Suhu dan Kalor," *J. Ris. Pendidik. Fis.*, vol. 6, no. 1, pp. 20–26, 2021.

[5] S. M. Ni'mah, "Analisis miskonsepsi dan kemampuan berpikir kritis siswa pada materi suhu dan kalor menggunakan pembelajaran modeling instruction disertai web-based formative assessment." Universitas Negeri Malang, 2019.

[6] R. Priyadi, K. Suryanti, and L. Varela, "Profil model pemahaman peserta didik pada topik suhu dan kalor: Studi lintas pendidikan," *J. Penelit. Pembelajaran Fis.*, vol. 10, no. 1, pp. 51–56, 2019.

[7] I. Y. Okyranida, S. Mayanty, and F. Widiyatun, "Analisis Butir Soal Kemampuan Berpikir Kritis Siswa SMAIT Nururrohmah Depok," *J. Penelit. Pembelajaran Fis.*, vol. 15, no. 1, pp. 73–79, 2024, doi: 10.26877/jp2f.v15i1.17057.

[8] D. Wahyudin *et al.*, *Kajian Akademik Kurikulum Merdeka*. 2024.

[9] P. Kissi, D. Baidoo-Anu, E. Anane, and R. K. Annan-Brew, "Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence," in *Frontiers in Education*, Frontiers Media SA, 2023, p. 1154592.

[10] K. Ningsih and Afandi, "SKILLS ASSESSMENT BOOK FOR THE SUBJECT OF BIOLOGY IN GRADE VIII," *J. Mat. dan IPA*, vol. 15, no. 3, pp. 321–330, 2024.

[11] A. H. Pranata and I. Dwijayanti, "Analisis Kesulitan Penyusunan Instrumen Penilaian Bagi Guru Sekolah Dasar Di Gugus Dewi Sartika Kecamatan Reban," *J. Pendidik. Sekol. Dasar*, vol. 3, no. 2, pp. 61–74, 2021.

[12] D. Dziob, "Board game in physics classes—A proposal for a new method of student assessment," *Res. Sci. Educ.*, vol. 50, no. 3, pp. 845–862, 2020.

[13] S. H. P. W. Gamage, J. R. Ayres, and M. B. Behrend, "A systematic review on trends in using Moodle for teaching and learning," *Int. J. STEM Educ.*, vol. 9, no. 1, p. 9, 2022.

[14] V. Mardian, A. Samsudin, J. A. Utama, and I. R. Suwarma, "Validity and Reliability of Elasticity Multiple-Choice Items (EMCI) Using Rasch Model," vol. 12, no. 2, pp. 265–276, 2023, doi: 10.24042/jipfalbiruni.v12i2.17037.

[15] K. D. Rainey, M. Vignal, and B. R. Wilcox, "Designing upper-division thermal physics assessment items informed by faculty perspectives of key content coverage," *Phys. Rev. Phys. Educ. Res.*, vol. 16, no. 2, p. 20113, 2020, doi: 10.1103/PhysRevPhysEducRes.16.020113.

[16] S. Amalia, Y. Hendawati, and F.

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

Nuraeni, "Analisis Butir Soal Kemampuan Berpikir Kritis Siswa Sekolah Dasar," in *Renjana Pendidikan: Prosiding Seminar Nasional Pendidikan Dasar*, 2021, pp. 566–573.

[17] M. A. Henry, S. Shorter, L. K. Charkoudian, J. M. Heemstra, B. Le, and L. A. Corwin, "Quantifying fear of failure in STEM: Modifying and evaluating the Performance Failure Appraisal Inventory (PFAI) for use with STEM undergraduates," *Int. J. STEM Educ.*, vol. 8, pp. 1–28, 2021.

[18] M. Larrain and G. Kaiser, "Interpretation of students' errors as part of the diagnostic competence of pre-service primary school teachers," *J. für Math.*, vol. 43, no. 1, pp. 39–66, 2022.

[19] M. Pedaste, A. Baucal, and E. Reisenbuk, "Towards a science inquiry test in primary education: development of items and scales," *Int. J. STEM Educ.*, vol. 8, pp. 1–19, 2021.

[20] M. H. Towns, "Guide to developing high-quality, reliable, and valid multiple-choice assessments," *J. Chem. Educ.*, vol. 91, no. 9, pp. 1426–1431, 2014, doi: 10.1021/ed500076x.

[21] L. Ding and R. Beichner, "Approaches to data analysis of multiple-choice questions," *Phys. Rev. Spec. Top. - Phys. Educ. Res.*, vol. 5, no. 2, pp. 1–17, 2009, doi: 10.1103/PhysRevSTPER.5.020103.

[22] B. R. Rush, D. C. Rankin, and B. J. White, "The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value," *BMC Med. Educ.*, vol. 16, no. 1, p. 250, 2016.

[23] T. M. Haladyna, S. M. Downing, and M. C. Rodriguez, "A review of multiple-choice item-writing guidelines for classroom assessment," *Appl. Meas. Educ.*, vol. 15, no. 3, pp. 309–333, 2002.

[24] M. A. Bujang, E. D. Omar, and N. A. Baharum, "A review on sample size determination for cronbach's alpha test: A simple guide for researchers," *Malaysian J. Med. Sci.*, vol. 25, no. 6, pp. 85–99, 2018, doi: 10.21315/mjms2018.25.6.9.

[25] T. Winarsunu, *Statistik dalam penelitian psikologi dan pendidikan*, vol. 1. UMMPress, 2002.

[26] S. Arikunto, "Pendekatan Penelitian," *Jakarta: Rineka Cipta*, 1998.

[27] M. C. Ramadhani, S. Kantun, and J. Widodo, "Analisis Validitas dan Tingkat Kesukaran Soal Latihan Evaluasi Akhir Tahun Pada Buku Sekolah Elektronik (BSE) Mata Pelajaran Ekonomi SMA/MA Kelas XI (Studi Kasus pada SMA Negeri Rambipuji Jember Kelas XI IPS 1 Tahun Ajaran 2013-2014)," 2014.

[28] S. Nurlatifah, M. Supriatna, E. S. Yudha, and A. Julius, "Pengembangan Instrumen Kecakapan Kerja Mahasiswa

Jurnal Pendidikan Matematika dan IPA
Vol. 17, No. 2 (2026) h. 311-322

(Rasch Model Analisis). GUIDENA: Jurnal Ilmu Pendidikan, Psikologi," *Bimbing. Dan Konseling*, vol. 13, no. 4, p. 792, 2023.

[29] D. Adhari, Santiani, and S. Fatmawati, "VALIDITY AND RELIABILITY ANALYSIS OF AN ENVIRONMENTALLY INTEGRATED EduFisika : Jurnal Pendidikan Fisika Volume 10 Number 1 , April 2025," vol. 10, no. 1, pp. 38–53, 2025, doi: 10.59052/edufisika.v10i1.42136 .

[30] A. M. Farida and A. Musyarofah, "Validitas dan Reliabilitas dalam Analisis Butir Soal," *Al-Mu'arrib J. Pendidik. Bhs. Arab*, vol. 1, 2021.

[31] N. Z. Zuhri, S. Syihabuddin, and T. Tatang, "Analisis Validitas, Reliabilitas, dan Tingkat Kesukaran Soal Bahasa Arab Tingkat SMP Berbasis Artificial Intelligence (AI) melalui Platform QuestionWell," *J. Pendidik. dan Pembelajaran Indones.*, vol. 4, no. 2, pp. 693–704, 2024, doi: 10.53299/jppi.v4i2.576.

[32] G. F. Kuder and M. W. Richardson, "The theory of the estimation of test reliability," *Psychometrika*, vol. 2, no. 3, pp. 151–160, 1937.

[33] A. Tongchai, M. D. Sharma, I. D. Johnston, K. Arayathanitkul, and C. Soankwan, "Developing, evaluating and demonstrating the use of a conceptual survey in mechanical waves," *Int. J. Sci. Educ.*, vol. 31, no. 18, pp. 2437–2457, 2009, doi: 10.1080/09500690802389605.

[34] A. Lichtenberger, C. Wagner, S. I. Hofer, E. Stern, and A. Vaterlaus, "Validation and structural analysis of the kinematics concept test," *Phys. Rev. Phys. Educ. Res.*, vol. 13, no. 1, p. 10115, 2017.

[35] A. Jufriadi, "Instrumen kinematics concept inventory: pengembangan dan validasi." Universitas Negeri Malang, 2024

.