Journal of Mathematics Instruction, Social Research and Opinion
Vol. 5, No. 2, June 2026, pp. 1979 – 1988, https://doi.org/10.58421/misro.v5i2.1508
ISSN 2962-7842

Qualitative and Quantitative Item Analysis of Essay Test Instruments in a Learning Planning and Microteaching Course

Ranu Iskandar[1], Putri Khoirin Nashiroh[2], Muhammad Khumaedi[3]
[1,2,3]Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received 2026-04-28
Revised 2026-06-27
Accepted 2026-06-27

Keywords:

Item discrimination
Item quality analysis
Learning, planning, and microteaching
Level of difficulty
Reliability
Validity

ABSTRACT

This study aimed to conduct qualitative and quantitative analyses of essay test instruments used in the daily assessment of the Learning Planning and Microteaching course. This research employed a descriptive method with an item analysis approach. The subjects of this study were fifth-semester students of the Automotive Engineering Education Study Program, while the object of the study consisted of 15 essay items used in the daily assessment of the Learning Planning and Microteaching course. The study was conducted on September 25, 2025, involving 20 student respondents. Qualitative analysis of the test instrument was carried out by an expert to detect and correct weaknesses in the material, construction, and language aspects so that the instrument is valid in content and objective before being used. Quantitative analysis of test instruments is the process of testing the quality of test instrument items to determine item discrimination, Level of difficulty, and external validity reliability. The results showed that, based on the qualitative analysis, all items met the criteria for material, construction, and language aspects. However, the quantitative analysis, including difficulty index, discrimination index, validity, and reliability, indicated that several items required revision. Items 11 and 12 need revision because they showed weaknesses in validity, although their difficulty and discrimination indices were still acceptable. Meanwhile, items 2, 5, 8, 14, and 15 require more substantial revision or replacement because they have poor discrimination power. Overall, the essay test instrument can be considered moderately feasible, but it still requires improvement before being used as a strong formative assessment instrument.

Corresponding Author:

Ranu Iskandar
Automotive Engineering Education, Faculty of Engineering, Universitas Negeri Semarang, Indonesia
Email:ranuiskandar@mail.unnes.ac.id

1. INTRODUCTION

The Learning Planning and Microteaching course is offered in the fifth semester by the Automotive Engineering Education Study Program, Faculty of Engineering, Universitas Negeri Semarang [1]. The learning objective of this course is to train students to apply

various forms of knowledge, attitudes, and teaching skills through different teaching methods and learning models in order to prepare them to become professional teachers [2][3]. In this course, students learn about the nature of microteaching, the objectives and benefits of microteaching, the characteristics of microteaching, the procedures for implementing microteaching, general teaching procedures, learning planning, microteaching practice, teaching implementation, and basic teaching skills. To determine whether the Learning Planning and Microteaching course has achieved its learning objectives, an appropriate learning assessment is required [4][5].

Assessment is an important component in determining the success of the learning process conducted by teachers [6]. It also provides feedback to teachers so that they can continuously improve their teaching abilities and help students achieve optimal learning development [7]. Assessment conducted by teachers should not only measure learning outcomes but also become part of the effort to support students' learning process [8]. Teachers need to realise that students' learning success is one of the indicators of teachers' success in teaching [9]. Apart from that, the aim of micro-learning is for students to master 8 basic teaching skills, one of which is to close the learning by giving test questions [10].

Learning assessment can be conducted through achievement tests. Achievement tests may take the form of written tests, performance tests, or oral tests. Written tests may include multiple-choice items, essay items, and matching items [11]. An essay test consists of questions or tasks that require students to organise and express their answers in their own words or sentences. These answers may involve recalling, organising, integrating, or constructing knowledge that has been learned into coherent written responses. Essay tests not only measure low-order thinking skills (LOTS) but also higher-order thinking skills (HOTS) [12]. The development of essay items begins with several stages: (1) formulating the purpose of the test, (2) reviewing which learning materials are suitable for essay-type questions, (3) developing a test blueprint, (4) writing the questions along with answer keys and scoring guidelines, and (5) determining the quality of the test items [13]. The quality of test items can be determined through qualitative and quantitative analyses. Quantitative analysis includes difficulty index, discrimination index, validity, and reliability analysis.

The difficulty index refers to the proportion of students who are able to answer an item correctly compared to the total number of test takers. Difficulty index analysis is conducted to determine whether an item is categorised as easy, moderate, or difficult. In other words, the difficulty index is calculated based on the proportion of students who answer the item correctly. When more students are able to answer correctly, the difficulty index becomes higher, indicating that the item is easier. Conversely, when fewer students answer correctly, the item is categorised as difficult. The discrimination index refers to the ability of an item to distinguish between high-achieving and low-achieving students. Since essay test instruments need to meet acceptable assessment standards, their validity and reliability must also be demonstrated [14].

The purpose of this study was to qualitatively review essay items and quantitatively analyse the essay items in terms of difficulty index, discrimination index, validity, and reliability in the daily assessment of the Learning Planning and Microteaching course. The main novelty of this research is that it provides empirical evidence regarding the technical

quality of essay tests and offers a model for lecturers to improve formative assessment instruments. The novelty of this research is that the questions are arranged based on Bloom's Taxonomy, starting from level C1 to C3 (LOTS) and C4 to C5 (HOTS).

The research gap addressed in this study is that previous studies on test instrument analysis in education have mostly focused on objective tests, such as multiple-choice, true-false, or matching items. Meanwhile, essay test instruments have rarely been analysed in depth, particularly in terms of validity, reliability, discrimination power, and difficulty level.

2. METHOD

This study employed a quantitative method aimed at analysing the quality of essay items used in the daily assessment of the Learning Planning and Microteaching course. The subjects of this study were fifth-semester students of the Automotive Engineering Education Study Program, while the object of the study was the essay test used in the daily assessment of the Learning Planning and Microteaching course. The questions consisted of 4 questions at level C1 and C2 with a maximum score of 5, 6 questions at level C3 with a maximum score of 7, 2 questions at level C4 with a maximum score of 9, and 2 questions at level C5 with a maximum score of 10, so the total score is 100. The study was conducted on September 25, 2025, involving 20 student respondents. The sample size was only 20 respondents because I only taught 1 class consisting of 20 students. The data used in this study were primary data, collected directly from respondents through test instruments.

The quality of the test items was reviewed using the format developed by the Centre for Educational Assessment, Indonesia Ministry of Education and Culture [15]. An expert reviewed the essay test items as a lecturer. The aspects assessed included material, construction, and language.

The item difficulty index refers to the probability of answering an item correctly at a certain level of ability and is commonly expressed as an index. The formula for the item difficulty index is as follows [16]:

$$DIF = [(H+L)/N] \times 100 \quad (1)$$

Description
DIF = difficulty Index
H = number of students who gave correct options in the high-score group
L = number of students who gave correct options in the low-score group
N = Total number of students who took the test

Table 1. Difficulty Index

No	Category	Range
1	Too easy	>70%
2	Good	30-70%
3	Too difficult	<30%

The discrimination index is generally expressed as a proportion ranging from 0 to 1.00. It indicates how well an item can distinguish between students who have mastered the

material and those who have not. The higher the discrimination index, the better the item is in measuring students' mastery of the material. The formula for the discrimination index is as follows [17].

$$DI = (H - L) \times 2/N \quad (2)$$

Description

DI = Discrimination index
H = number of students answering the item correctly in the high-achieving group
L = number of students answering the item correctly in the low-achieving group
N = total number of students in the two groups (including non-responders)

Table 2. Discrimination Index Analysis

No	Category	Range
1	Poor	0.00-0.20
2	Enough	0.20-0.40
3	Good	0.40-0.70
4	Very Good	0.70-1.00

Construct validity in the test instrument was examined using Pearson's Product-Moment correlation, particularly to determine the relationship between each item score and the total score of the construct or variable [18]. An item is considered valid when the calculated r-value is higher than the r-table value [19]. The formula for Pearson's Product-Moment correlation is as follows [20]:

$$r_{xy}=\frac{\mathbf{N}\sum\mathbf{XY}-(\sum\mathbf{X})(\sum\mathbf{Y})}{\sqrt{[\mathbf{N}\sum\mathbf{X}^2-(\sum\mathbf{X})^2]\,[\mathbf{N}\sum\mathbf{Y}^2-(\sum\mathbf{Y})^2}} \quad (3)$$

Description

r_{xy} = Product-moment correlation coefficient
N = number of respondents
X = score of each item
Y = Total score

The reliability estimate in this study was examined using Cronbach's Alpha. The Cronbach's Alpha coefficient indicates the extent to which the items in one variable or construct are internally consistent in measuring the same concept [21]. An instrument is considered reliable when the Cronbach's Alpha value is greater than 0.60 [22]. The formula for Cronbach's Alpha is as follows [23]:

$$\alpha=\frac{\mathbf{k}}{\mathbf{k-1}}\ (1-\sigma_t^2\textstyle\sum\sigma_i^2) \quad (4)$$

Description

α = Cronbach's Alpha reliability coefficient
k = number of items
σi^2 = Variance of each item
σt^2 = Variance of the total score

3. RESULTS AND DISCUSSION

The quality of a test instrument is essential in learning evaluation to ensure that the questions given are able to measure students' abilities accurately. In this study, the quality of the test instrument was analysed both qualitatively and quantitatively. The qualitative analysis was conducted by an expert based on material, construction, and language aspects, while the quantitative analysis included difficulty index, discrimination index, validity, and reliability.

Table 3. Analysis of the Quality of the Essay Test for Microteaching Courses

No	Aspect Reviewed	Item Number														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Material															
1	The item is aligned with the indicator.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	The scope of the question and the expected answer are appropriate.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	The material tested is aligned with the required competency.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	The content of the material tested is appropriate to the school level or grade level.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
B.	Construction															
5	Tables, figures, graphs, maps, or similar elements are presented clearly.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	The item uses question words or commands that require an essay-type response.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
7	Clear instructions are provided for answering the item.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C	Bahasa															
8	The wording of the item is communicative.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
9	The item uses standard Indonesian language.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	The item does not use words or expressions that may cause multiple interpretations or misunderstanding.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
11	The item does not use local, colloquial, or taboo language.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
12	The wording of the item does not contain words or expressions that may offend students.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Recommendation	Accepted/~~Revised/Rejected~~ *)														

In this study, the qualitative analysis showed that all essay items had met the criteria for the material, construction, and language aspects. This finding indicates that, first, each item was aligned with the learning materials and learning objectives. Second, each item was able to measure the intended construct or variable. Lastly, the language used in each item was easy for respondents to understand. The sentences were not ambiguous, overly long, difficult in terminology, or double-meaning, thereby minimising the possibility of respondents misunderstanding the intent of the questions [24]. However, qualitative analysis does not automatically guarantee that all items function well empirically; therefore, quantitative item analysis remains necessary [25].

Table 4. Difficulty Index Analysis

No	Item Number	Amount	Percentage	Interpretation
1	5, 8, 14 ,15	4	26%	Too difficult
2	1, 3, 4, 6, 7, 9, 10, 11, 12, 13	10	67%	Good
3	2	1	7%	Too difficult

The difficulty index showed that the items were distributed into several categories: too easy (>70%), acceptable (30-70%), and too difficult (<30%) [26]. Items 5, 8, 14, and 15 were categorised as too easy; items 1, 3, 4, 6, 7, 9, 10, 11, 12, and 13 were categorised as acceptable; and item 2 was categorised as too difficult. This distribution indicates that the test had a variety of difficulty levels, but the proportion of too-easy items was still relatively high. Items that are too easy may be useful for measuring basic understanding, but they are less effective for identifying students with high analytical ability [27]. In the context of microteaching, essay items should not only ask students to recall concepts but also require them to analyse teaching scenarios, design learning activities, justify teaching strategies, and reflect on classroom practice [28][29].

Table 5. Discrimination Index Analysis

No	Item Number	Amount	Percentage	
1	2, 5, 8, 14, 15	6	0.40	Enough
2	1, 3, 7, 9, 11, 13	5	0.33	Enough
3	4, 6, 10, 12	4	0.27	Enough

The discrimination index showed that the items were distributed into several categories: poor, enough, and good [30]. The discrimination index revealed that items 4, 6, 10, and 12 had good discrimination power, while items 1, 3, 7, 9, 11, and 13 were in the sufficient category. These items can still be used because they were able to differentiate, to a certain extent, students with higher and lower mastery [31]. However, items 2, 5, 8, 14, and 15 were categorised as poor. These items failed to distinguish students with high ability from those with low ability. Items with poor discrimination should therefore be reviewed because a low discrimination index may indicate that the item is not functioning well in differentiating students based on their mastery level [32]. Meanwhile, item 2 was categorised as difficult and also had poor discrimination, indicating a possibility of ambiguity in the question, misalignment with learning materials, or an overly strict scoring criterion [33].

Table 6. Construct Validity Test Using Pearson's Product-Moment Correlation

	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
rcount	0.449	0.485	0.582	0.483	0.450	0.607	0.400	0.183	0.477	0.606	0.378	0.247	0.471	0.513	0.142
rtab								0,444							
sig	0.047	0.030	0.007	0.031	0.036	0.005	0.081	0.439	0.045	0.005	0.100	0.293	0.036	0.011	0.551
α								0,05							
df								20-2=18							
Status	V	V	V	V	V	V	N	N	V	V	N	N	V	V	N

The validity test using the Pearson Product-Moment correlation showed that 10 out of 15 items were valid, namely items 1, 2, 3, 4, 5, 6, 9, 10, 13, and 14. These items had correlation coefficients higher than the r-table value of 0.444 or significance values below 0.05. Meanwhile, items 7, 8, 11, 12, and 15 were categorised as invalid. This finding indicates that most items were able to represent the total construct measured by the test; however, one-third of the items did not show a strong correlation with the total score. Validity in educational assessment is not merely a statistical property of the instrument, but evidence that supports the interpretation of test scores in relation to the intended construct and learning objectives [34]. Therefore, invalid essay items may be caused by several factors, such as weak alignment between the item and the measured indicator, overly broad expected responses, insufficiently specific scoring rubrics, or a mismatch between the cognitive demand of the item and the construct being assessed. Previous assessment studies emphasise that test items should be aligned with specific content, instructional objectives, and intended cognitive processes, while constructed-response tasks require well-defined scoring procedures and rubric interpretation to reduce construct-irrelevant score variation [35][36]. Therefore, items 7, 8, 11, 12, and 15 need to be revised or removed before being reused in further assessment.

The reliability analysis using Cronbach's Alpha produced a coefficient of 0.567. This value indicates that the internal consistency of the essay test was in the sufficient category but had not yet reached a strong level, with a minimum of 0.70 [37]. This condition shows that the items were not fully consistent in measuring students' mastery of learning planning and microteaching concepts. The relatively moderate reliability may be influenced by the small number of respondents [38].

Table 7. Reliabilty

Cronbach's Alpha	N of Items	Interpretation
0.567	15	Poor reliability

Based on the combination of validity, difficulty index, and discrimination index, items 3, 4, 9, 10, and 13 can be considered the strongest items because they were valid, had moderate difficulty, and had sufficient or good discrimination power. Items 1 and 6 may still be used with minor revision because they were valid and had acceptable discrimination, although their difficulty levels were relatively easy. Items 7, 11, and 12 need revision because they showed weaknesses in validity, even though their difficulty or discrimination indices were still acceptable [39]. Items 2, 5, 8, 14, and 15 require more substantial revision or replacement because they have poor discrimination power [40].

Overall, the essay test instrument for the learning planning and microteaching course can be considered moderately feasible, but it still requires improvement before being used as a strong formative assessment instrument. The findings provide empirical evidence that essay tests in microteaching need systematic item analysis because pedagogical competence and teaching readiness cannot be measured only by the availability of questions and rubrics. The revised items should be directed toward more authentic microteaching tasks, such as analysing lesson plans, selecting appropriate teaching models, formulating assessment

strategies, and reflecting on teaching performance. Thus, the instrument can better support the measurement of students' pedagogical understanding and practical teaching readiness.

In addition, the interpretation of the results should consider the limitations of the classical test theory approach, in which the group of test participants influences item characteristics. Since this study involved 20 students, the results need to be interpreted carefully and may change if the instrument is tested with a larger and more diverse sample. Further research is recommended to revalidate the revised items, involve more respondents, strengthen inter-rater reliability in essay scoring, and compare the results with other analytical approaches. These steps will help produce a more valid, reliable, and practically useful essay test instrument for microteaching assessment.

This study has several limitations. First, the sample size was small, involving only 20 students, so the item statistics should be interpreted carefully. Second, the study was conducted only in one course, namely Learning Planning and Microteaching, which limits the applicability of the findings to other courses or study programs. Third, the qualitative review involved only one expert reviewer, so the content judgment may not fully represent broader expert perspectives. Fourth, this study did not examine inter-rater reliability, even though essay scoring is highly dependent on scorer consistency. Fifth, the findings have limited generalizability because the data were collected from a specific class context. Sixth, the analysis used only classical test theory, in which item characteristics may depend on the group of test takers. Future studies should involve larger samples, more expert reviewers, inter-rater reliability analysis, pilot testing of revised items, and additional analytical approaches to strengthen the evidence of item quality.

4. CONCLUSION

The essay test is acceptable but still needs improvement. Although all items passed the qualitative review, the quantitative results show that some items still have weak validity, poor discrimination, and low reliability. Therefore, the instrument should be revised before being used again. Items 3, 4, 9, 10, and 13 can be retained because they showed the strongest quality. Items 1 and 6 need minor revision, especially in wording and scoring criteria. Items 7, 11, and 12 need major revision because they have validity problems. Items 2, 5, 8, 14, and 15 should be replaced because they had poor discrimination power. The main improvement should focus on the scoring rubric. The rubric should include clearer answer criteria, score levels, and examples of expected responses. After revision, the items and rubric should be revalidated with more students and tested again for validity, reliability, and scoring consistency.

ACKNOWLEDGEMENTS

Thank you to all parties who have helped with this research.

REFERENCES

[1] Universitas Negeri Semarang, “Pendidikan Teknik Otomotif: Kurikulum.” 2025. [Online]. Available: https://unnes.ac.id/ft/id/pto-kurikulum/

[2] C. Yuliana *et al.*, *Microteaching: Strategi Microteaching dalam Pembelajaran Efektif*. KOta Jambi: PT. Sonpedia Publishing Indonesia, 2025.

[3] J. O’Flaherty, R. Lenihan, A. M. Young, and O. McCormack, “Developing Micro-Teaching with a

Focus on Core Practices: The Use of Approximations of Practice," *Educ. Sci.*, vol. 14, no. 1, p. 35, 2023, [Online]. Available: https://www.mdpi.com/2227-7102/14/1/35

[4] M. Barnard, E. Whitt, and S. McDonald, "Learning objectives and their effects on learning and assessment preparation: insights from an undergraduate psychology course," *Assess. Eval. High. Educ.*, vol. 46, no. 5, pp. 673–684, 2021, [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/02602938.2020.1822281

[5] R. Iskandar, "Assessing the Digital Literacy Profile of Productive Automotive Engineering Teacher Candidate Students," *J. Educ. Teach.*, vol. 5, no. 1, pp. 60–69, 2024, [Online]. Available: https://jet.or.id/index.php/jet/article/view/331

[6] I. Magdalena, H. N. Fauzi, and R. Putri, "Pentingnya evaluasi dalam pembelajaran dan akibat memanipulasinya," *Bintang*, vol. 2, no. 2, pp. 244–257, 2020, [Online]. Available: https://ejournal.stitpn.ac.id/index.php/bintang/article/view/986

[7] T. Siregar, *Micro Teaching*. Kabupaten Cirebon: Goresan Pena, 2025.

[8] R. Lasso, "A Blueprint for Using Assessments to Achieve Learning Outcomes and Improve Students' Learning," *SSRN*, 2020, [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3406301

[9] I. Đerić, I. Elezović, and F. Brese, "Teachers, Teaching and Student Achievement," in *Dinaric Perspectives on TIMSS 2019. IEA Research for Education*, B. Japelj Pavešić, P. Koršňáková, and S. Meinck, Eds., Cham: Springer, 2022, pp. 151–174. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-030-85802-5_7

[10] S. Suryana, *Pembelajaran Mikro*. Sukoharjio: Tahta Media Group, 2024.

[11] R. Iskandar, *Pedoman Penilaian Hasil Belajar Peserta Didik SMK Kompetensi Keahlian Teknik Kendaraan Ringan pada Mata Pelajaran Pemeliharaan Sasis Dan Pemindah Tenaga Kendaraan Ringan*. Sukabumi: CV Jejak (Jejak Publisher), 2019.

[12] C. Poluakan and A. L. . Tilaar, "Hots dan lots: realiti atau ilusi?," *J. Eval. Dan Pembelajaran*, vol. 2, no. 1, pp. 88–94, 2020, doi: 10.52647/jep.v2i1.16.

[13] Kementerian Pendidikan dan Kebudayaan, *Penilaian hasil belajar: Pendidikan dan pelatihan teknis kegiatan belajar mengajar bagi pamong belajar*. Jakarta: Kementerian Pendidikan & Kebudayaan, 2016. [Online]. Available: https://repositori.kemendikdasmen.go.id/17902/1/03.15 Modul Pelatihan TFM bagi Pamong Belajar 05. Penilaian Hasil Belajar.pdf

[14] B. Quah, L. Zheng, T. J. H. Sng, C. W. Yong, and I. Islam, "Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations," *BMC Med. Educ.*, vol. 24, no. 1, p. 962, 2024, [Online]. Available: https://link.springer.com/article/10.1186/s12909-024-05881-6

[15] Pusat Penilaian Pendidikan, *Panduan Penilaian Tes Tertulis*. Jakarta: Kementerian Pendidikan dan Kebudayaan, 2019. [Online]. Available: https://repositori.kemdikbud.go.id/18344/1/PANDUAN PENILAIAN TERTULIS 2019.pdf

[16] P. Kunjappagounder, S. K. Doddaiah, P. N. Basavanna, and D. Bhat, "Relationship between Difficulty and Discrimination Indices of Essay Questions in for Mative Assessment," *J. Anat. Soc. India*, vol. 70, no. 4, pp. 239–243, 2021, [Online]. Available: https://journals.lww.com/joai/fulltext/2021/70040/relationship_between_difficulty_and_discrimination.9.aspx

[17] A. Chauhan, F. Khaliq, and K. R. Nayak, "Assessing Quality of Scenario-Based Multiple-Choice Questions in Physiology: Faculty-Generated vs. ChatGPT-Generated Questions among Phase I Medical Students," *Int J Artif Intell Educ*, vol. 35, pp. 2315–2344, 2025, [Online]. Available: https://link.springer.com/article/10.1007/s40593-025-00471-z

[18] N. F. Adkha, P. Sudira, and R. Iskandar, "The mindfulness aspects in the teaching of culinary art in vocational high school," *J. Pendidik. Vokasi*, vol. 11, no. 2, pp. 155–170, 2021, [Online]. Available: https://journal.uny.ac.id/index.php/jpv/article/view/38402

[19] "The Influence of Social Media Usage on the Authority of Religious Leaders Among Bina Nusantara University Students, Alam Sutera, Tangerang," in *Proceedings of TEEM 2024*, Singapore: Springer, 2025, pp. 813–820. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-96-5658-5_80

[20] T. Gnambs, "A Brief Note on the Standard Error of the Pearson Correlation," *Collabra Psychol.*, vol. 9, no. 1, p. 87615, 2023, [Online]. Available: https://online.ucpress.edu/collabra/article/9/1/87615/197169

[21] C. G. Forero, "Cronbach's Alpha," *Encyclopedia of Quality of Life and Well-Being Research*, 2024. https://link.springer.com/rwe/10.1007/978-3-031-17299-1_622

[22] X. Zhang *et al.*, "Reliability and Validity of the Tilburg Frailty Indicator in 5 European Countries," *J. Am. Med. Dir. Assoc.*, vol. 21, no. 6, pp. 772-779.e6, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1525861020302784

[23] I. Bin Sa'id *et al.*, *KONSEP PENELITIAN KUANTITATIF*. Kota Padang: CV. PUSTAKA INSPIRASI MINANG, 2024. [Online]. Available: https://opac.upgripnk.ac.id/index.php?p=fstream-pdf&fid=190&bid=8442

[24] A. R. Artino, J. S. La Rochelle, K. J. Dezee, and H. Gehlbach, "Developing questionnaires for educational research: AMEE Guide No. 87," *Med. Teach.*, vol. 36, no. 6, pp. 463–474, 2014, [Online]. Available: https://www.tandfonline.com/doi/full/10.3109/0142159X.2014.889814

[25] D. A. Cook and T. J. Beckman, "Current concepts in validity and reliability for psychometric instruments: theory and application," *Am J Med*, vol. 119, no. 2, pp. 166.e7–16, 2006, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0002934305010375

[26] W. Mahjabeen *et al.*, "Difficulty Index, Discrimination Index and Distractor Efficiency in Multiple Choice Questions," *Ann. PIMS*, vol. 13, no. 4, pp. 310–315, 2017, [Online]. Available: https://www.apims.net/apims/article/view/9/

[27] G. T. L. Brown, E. R. Peterson, and E. S. Yao, "Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades," *Front. Educ.*, vol. 2, p. 24, 2017, [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2017.00024/full

[28] M. Karlström and K. Hamza, "Preservice science teachers' opportunities for learning through reflection when planning a microteaching unit," *J. Sci. Teacher Educ.*, vol. 30, no. 1, pp. 44–62, 2019, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/1046560X.2018.1531345

[29] S. Ledger and J. Fischetti, "Micro-teaching 2.0: Technology as the classroom," *Australas. J. Educ. Technol.*, vol. 36, no. 1, pp. 37–54, 2020, [Online]. Available: https://ajet.org.au/index.php/AJET/article/view/4561

[30] Khairunnisa, A. H. Pulungan, and R. Husein, "Validity And Reliability Of The English Summative Test For Second Semester Of The Fifth Grade In Academic Year 2019/2020," *Int. J. Educ. Res. Soc. Sci.*, vol. 2, no. 1, pp. 92–101, 2021, [Online]. Available: https://ijersc.org/index.php/go/article/view/21/

[31] K. Quaigrain and A. K. Arhin, "Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation," *Cogent Educ.*, vol. 4, no. 1, p. 1301013, 2017, [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/2331186X.2017.1301013

[32] M. Tavakol and R. Dennick, "Post-examination analysis of objective tests," *Med. Teach.*, vol. 33, no. 6, pp. 447–458, 2011, [Online]. Available: https://www.tandfonline.com/doi/full/10.3109/0142159X.2011.564682

[33] S. Sabri, "Item analysis of student comprehensive test for research in teaching beginner string ensemble using model-based teaching among music students in public universities," *Int. J. Educ. Res.*, vol. 1, no. 12, pp. 1–14, 2013, [Online]. Available: https://www.ijern.com/journal/December-2013/28.pdf

[34] S. M. Downing, "Validity: On meaningful interpretation of assessment data," *Med. Educ.*, vol. 37, no. 9, pp. 830–837, 2003, [Online]. Available: https://asmepublications.onlinelibrary.wiley.com/doi/10.1046/j.1365-2923.2003.01594.x

[35] D. F. McCaffrey, J. M. Casabianca, K. L. Ricker-Pedley, R. R. Lawless, and C. Wendler, *Best Practices for Constructed-Response Scoring*. Princeton: ETS, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12358

[36] J. Trace, V. Meier, and G. Janssen, "'I can see that': Developing shared rubric category interpretations through score negotiation," *Assess. Writ.*, vol. 30, pp. 32–43, 2016, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1075293516300435

[37] Y. D. Sangary, M. H. Asayesh, A. Asgharzadeh, and Z. Naghsh, "Psychometric of the Ferrer-Urbina multidimensional scale of sexual self concept (MSSSC) in the Iranian population," *BMC Psychol.*, vol. 14, p. 229, 2026, [Online]. Available: https://link.springer.com/article/10.1186/s40359-025-03883-7

[38] D. G. Bonett, "Sample Size Requirements for Testing and Estimating Coefficient Alpha," *J. Educ. Behav. Stat.*, vol. 27, no. 4, pp. 335–340, 2022, [Online]. Available: https://journals.sagepub.com/doi/10.3102/10769986027004335

[39] S. N. Ikhsaniyah, A. D. Kurnia, M. Zuroida, V. Pratiwi, and L. Hakim, "Analisis butir soal perpajakan pph pasal 21 menggunakan software anates pada pendekatan teori tes klasik," *PEKA*, vol. 12, no. 2, pp. 77–88, 2024, [Online]. Available: https://journal.uir.ac.id/index.php/Peka/article/view/19917

[40] O. R. Sabela, D. Krisdayanty, A. Z. Taqqiyah, L. Hakim, and V. Pratiwi, "Analisis Butir Soal HOTS Elemen Dokumen Berbasis Digital (FASE E) Menggunakan Program Anates," *Educ. Achievment J. Sci. Res.*, vol. 6, no. 1, pp. 251–262, 2025, [Online]. Available: https://pusdikrapublishing.com/index.php/jsr/article/view/2328