TRANSPUBLIKA
Precise. Resilience. Felicitous.

Review of Multidisciplinary Education, Culture and Pedagogy
https://ojs.transpublika.com/index.php/ROMEO
Online ISSN 2810-0743

https://doi.org/10.55047/romeo.v4i4.1996

Original Article

Analysis of the Quality of Tryout Questions in the Marketing Education Program Based on Pedagogical Content Knowledge for UKPPPG Readiness

Rachmad Hidayat[1*], Wening Patmi Rahayu[2]

[1,2]Universitas Negeri Malang, Malang, Indonesia
Email: [1)] rachmad.hidayat.fe@um.ac.id, [2)] wening.patmi.fe@um.ac.id

Received : 17 October - 2025
Accepted : 19 November - 2025
Published online : 21 November - 2025

Abstract

A fundamental issue in the educational ecosystem is the assessment of learning outcomes. In preparing teachers for the Teacher Professional Education Competency Test (UKPPPG), tryout tests serve as strategic instruments. For the Marketing Education Program, these test items must reflect required pedagogical and professional competencies, particularly Pedagogical Content Knowledge (PCK), which integrates content mastery with teaching strategies. This study aims to analyze the quality of tryout test items based on PCK in preparing Marketing Education Program teachers for the UKPPPG. Using a quantitative-descriptive approach with 108 participants (36 male; 72 female), data were analyzed with SPSS 31, focusing on item validity (point-biserial correlation), reliability (Cronbach's α), discrimination index (D), difficulty level (p), and distractor effectiveness. Results indicate sufficient reliability (α = 0.760), though item quality varied: some items were valid, marginal, or invalid—particularly weak in early indicators. Discrimination indices were mostly "fair," with several items rated "good" (D ≥ 0.40) suitable for retention, while "poor" items (D < 0.20) require revision or replacement. The difficulty distribution was unbalanced for summative testing (easy 34%, medium 37%, difficult 29%), suggesting dominance of easy items that reduce discrimination. Distractor analysis revealed an average of 2–3 functioning distractors per item (66%), though some were implausible and required revision. The implications highlight the need for systematic selection and revision of items (stem, key, and distractors), rebalancing difficulty levels, and repeated pilot testing to ensure the instrument achieves higher validity, reliability, and representativeness of PCK-based teacher competencies in 21st-century marketing education.

Keywords: Distractor Effectiveness, Discrimination Index, Difficulty Level, Teacher Professional Education Competency Test (UKPPPG), Tryout.

1. Introduction

A fundamental issue in the educational ecosystem is the assessment of learning outcomes. Assessment in the learning process functions as an indicator of the level of competency achievement attained by students. The implementation of learning assessment needs to be supported by thorough preparation. One preparatory step that can be taken to face examinations is conducting tryouts. The form of this measurement is a trial, and in the educational context, this trial is known as a test. Tryouts aim to prepare students in solving exam questions. Thus, tryouts can function as a pretest to assess basic abilities before taking the exam.

Tryouts become a strategic instrument in preparing teachers to face the Teacher Professional Education Competency Test (UKPPPG). Quality test items will help prospective teachers measure their abilities, improve weaknesses, and prepare learning strategies. In the

context of the Marketing study program at Vocational High Schools (SMK), tryout test items must reflect the complexity of pedagogical, professional, social, and personality competencies needed by marketing teachers. This aligns with the demands of the Merdeka curriculum that emphasizes mastery of Pedagogical Content Knowledge (PCK). Therefore, analysis of tryout test item quality becomes an important step to ensure that the tested items can represent the actual teacher competency standards.

Pedagogical Content Knowledge (PCK) as proposed by Shulman (1986) is a knowledge framework that connects teachers' understanding of content (content knowledge) with pedagogical delivery strategies (pedagogical knowledge). In marketing learning, teachers are required not only to master marketing theory but also to teach it contextually according to the characteristics of SMK students. Therefore, the quality of PCK-based tryout test items greatly determines teacher readiness in facing the UKPPPG. Analysis of test item quality in the context of educational evaluation refers to efforts to identify validity, reliability, difficulty level, discrimination power, and distractor effectiveness. Thus, test items not only function as measurement tools but also as diagnostic instruments that provide an overview of teachers' abilities in applying PCK.

However, the main problem often encountered is the lack of in-depth evaluation of tryout test item quality. Many questions are merely recalling knowledge and have not touched analytical and applicative dimensions. Seeing these problems, analysis of PCK-based tryout test item quality becomes very important. This analysis not only functions to improve item quality but also as a reflection on teacher readiness in developing meaningful learning. Thus, this research is expected to provide real contributions in improving the quality of assessment instruments for prospective teachers, especially in the Marketing Study Program. PCK-based tryout questions can facilitate teachers to understand how marketing concepts are taught applicatively, critically, and relevantly to the needs of the business and industrial world.

Prior scholarship has established foundational knowledge, Iskandar & Rizal (2018) and Anggreini & Darmawan (2017) provide methodologies for tryout item analysis, while Dewi et al. (2020) and Adipat et al. (2023) validate the Pedagogical Content Knowledge (PCK) framework for teaching. These research streams, however, remain unconnected. Existing item analyses do not utilize a PCK lens to assess whether questions measure the integration of pedagogy and content; similarly, PCK research does not emphasize framework application for evaluating tryout quality. This produces an unaddressed gap concerning the empirical analysis of whether tryouts for prospective Marketing teachers effectively evaluate the applied PCK competencies required by contemporary curricula which is the central focus of this investigation.

This research also enriches the scientific repertoire related to teacher education assessment. The focus on PCK as an analytical framework makes this research relevant to the development of modern educational theory that emphasizes the integration of pedagogy and content. This research also supports the policy direction of the Ministry of Basic and Secondary Education that emphasizes competency-based teacher professionalism. Therefore, based on this description, this research aims to analyze the quality of Marketing Study Program tryout test items based on PCK in order to prepare teachers to face the UKPPPG. The analysis results are expected to provide practical recommendations for item developers, professional education lecturers, and teacher education institutions in compiling assessment instruments that are more valid, reliable, and representative of 21st-century teacher competencies.

2. Methods

The research method used in this study is quantitative research with a descriptive approach. Research subjects consisted of 108 teachers participating in UKPPPG BGT Phase 1 Marketing Study Program who took the tryout. Research data were obtained from participants' answers to the prepared tryout questions. The focus of analysis is directed to describe the quality of test instruments through testing item validity, reliability, discrimination power, difficulty level, and distractor effectiveness. The descriptive approach was chosen because this research describes the quality of evaluation instruments used in tryouts in detail and factually.

Data analysis was conducted using SPSS version 31. Validity testing was done by calculating the correlation between item scores and total scores, while instrument reliability was tested using Cronbach's Alpha coefficient technique. Question discrimination power was calculated to see the extent to which each item can distinguish participants with high and low abilities, while difficulty level was analyzed based on the proportion of participants who answered each item correctly. In addition, distractor effectiveness was examined to assess the extent to which distractors in multiple-choice questions function according to their purpose. The results of this analysis are expected to provide a comprehensive picture of the quality of the tested instrument so that it can be used for improvement and refinement in subsequent evaluation implementation.

3. Results and Discussion

3.1. Research Results

Tryout participants were students of the Marketing Study Program Teacher Professional Education for Certain Teachers Phase One in 2025. There were 36 male test participants and 72 female participants. Test participants had diverse backgrounds including 2 people as PAUD teachers, 9 elementary school teachers, 1 special education school teacher, 7 junior high school teachers, 2 senior high school teachers, and 87 vocational high school teachers. Therefore, it can be concluded that only 87 people or 80% have backgrounds or fields that match the tested study program. Tryout results data can be seen in the following table:

Table 1. Results of Test Item Quality Analysis

No	Indicator	Validity		Reliability	Discrimination Power		Difficulty Level	
		Validity	Criteria		D	Criteria	p	Criteria
1	Mastering and applying student learning management with student-centered learning methods	0,335	Valid	0,760	0,345	Fair	0,870	Easy
2		0,057	Invalid		0,207	Poor	0,352	Medium
3		0,179	Invalid		0,310	Fair	0,759	Easy
4		0,165	Invalid		0,276	Poor	0,750	Easy
5		0,436	Valid		0,552	Fair	0,796	Easy
6		0,346	Valid		0,517	Fair	0,713	Easy
7		-0,172	Invalid		-0,138	Poor	0,093	Difficult
8		0,171	Invalid		0,103	Poor	0,065	Difficult
9		0,541	Valid		0,724	Good	0,417	Medium
10		-0,145	Invalid		0,069	Poor	0,667	Medium
11		0,498	Valid		0,655	Fair	0,287	Difficult
12	Stable personality ability, noble	0,219	Invalid		0,276	Poor	0,167	Difficult
13		0,428	Valid		0,552	Fair	0,241	Difficult
14		0,383	Valid		0,552	Fair	0,685	Medium

15	character, wise, and authoritative as well as being a role model for students. This personality ability is carried out through reflection in carrying out responsibilities as a teacher according to the professional code of ethics and oriented toward students	0,161	Invalid	0,241	Poor	0,389	Medium
16		-0,019	Invalid	0,000	Poor	0,722	Easy
17		0,152	Invalid	0,207	Poor	0,769	Easy
18	Teachers' ability to communicate and interact effectively and efficiently with students, fellow teachers, parents, student guardians, and the surrounding community	0,332	Valid	0,138	Poor	0,963	Easy
19		0,140	Invalid	0,241	Poor	0,352	Medium
20		0,083	Invalid	0,138	Poor	0,222	Difficult
21		0,451	Valid	0,655	Fair	0,500	Medium
22		0,335	Valid	0,276	Poor	0,907	Easy
23		0,199	Invalid	0,345	Fair	0,398	Medium
24	Mastering subject matter broadly and deeply to determine learning objectives and organize knowledge content of student-centered learning	0,276	Invalid	0,379	Fair	0,231	Difficult
25		0,311	Valid	0,621	Fair	0,574	Medium
26		0,330	Valid	0,448	Fair	0,269	Difficult
27		0,436	Valid	0,655	Fair	0,472	Medium
28		0,283	Invalid	0,379	Fair	0,778	Easy
29		0,324	Valid	0,379	Fair	0,333	Medium
30		0,126	Invalid	0,138	Poor	0,741	Easy
31		0,512	Valid	0,690	Fair	0,287	Difficult
32		0,069	Invalid	0,172	Poor	0,454	Medium
33		0,371	Valid	0,483	Fair	0,694	Medium
34		0,107	Invalid	0,207	Poor	0,880	Easy
35		0,416	Valid	0,483	Fair	0,204	Difficult

Table 2. Summary of Analysis Results

Question Type	Validity Level	Reliability Level	Discrimination Power	Difficulty Level	Distractor Function
Multiple choice	Low-medium	High	High 3%, Ideal 51%, Low 46%	Difficult 29%, Medium 37%, Easy 34%	Functioning 28%, Partial 52%, Non-functioning 20%

3.1.1. Item Validity

Item validity analysis results using point-biserial correlation (rpB) on a tryout containing 35 items show a mixed composition between valid and invalid items. Referring to general standards, items with rpB ≥ 0.30 are categorized as valid, 0.20-0.29 are marginal, and < 0.20 or negative are considered invalid. In Indicator 1 block (initial items), many items are marked invalid even some have small or negative values such as items number 2, 3, 4, as well as 7, 8, 10. Conversely, a number of items appear valid with medium, high rpB, including numbers 1, 5, 6, 11, 20, 25, 27, 33, and 34. This pattern suggests that construct representation in initial indicators is still weak, while later indicators tend to be more consistent. The implication is that items that are already valid should be retained for the final test composition. Marginal items (rpB ≈ 0.20–0.29) should be revised by clarifying stimulus/wording, balancing distractor strength, and avoiding term ambiguity. Items with low or negative rpB should be dropped or replaced, especially if aligned with findings of poor discrimination power (D) and/or extremely difficult level (p). After improvements are made, retesting should be conducted to confirm improvements in instrument validity and reliability while reviewing difficulty distribution and discrimination power so that the instrument is more representative of the measured construct.

3.1.2. Reliability

Reliability analysis results show a Cronbach's α value of 0.760, which falls into the fair to good category for the instrument trial phase. This indicates that overall, the test items have adequate internal consistency in measuring the same construct. According to Tavakol & Dennick (2011), α values between 0.70–0.80 are commonly accepted in the early phase of test development, with opportunities for improvement if invalid or low discrimination items are deleted, and the composition of difficulty levels (easy-medium-difficult) is balanced. Practically, item selection strategies can be done with two approaches: top-down, which is deleting the weakest items so that reliability increases, or bottom-up, which is retaining items with the best correlations to strengthen the instrument. Thus, the value α = 0.760 shows a fairly good foundation, but can still be improved through item selection and revision to achieve a more reliable instrument.

3.1.3. Discrimination Power

Discrimination power (D) analysis results show that items' ability to distinguish high and low ability participants still varies: the majority of items are in the fair category (0.20-0.39), there are a number of poor items (<0.20) such as no. 2, 4, 8, 13, 16, 23, 29, 31-32, and there are also good items (≥0.40) such as no. 5, 6, 11, 20, 25, 27, 33. Operationally within the Classical Test Theory (CTT) framework (e.g., using point-biserial/RIT), negative or very low D values indicate serious problems with items (wording/key/distractor/interpretation), while higher D indicates items are more representative of the construct because their responses align with differences in participant abilities. The implication is to retain items with D ≥ 0.40 for the final item bank, revise items with D 0.20-0.39 by checking stem clarity, option length/structure equivalence, potential for overly prominent key patterns; and replace items with D < 0.20, especially if also invalid.

3.1.4. Difficulty Level

Difficulty level analysis results of 35 tryout test items show that item distribution is still not balanced according to ideal composition. Many items fall into the easy category (p > 0.70), such as items with scores of 0.87; 0.75; 0.79; 0.96; 0.91; up to 0.88, which dominate the item group. Meanwhile, there are a number of items in the medium category (0.30-0.70), for

example items with scores of 0.35; 0.41; 0.50; 0.57; 0.69, which are relatively sufficient to provide variation. On the other hand, there are also several items classified as difficult ($p < 0.30$), such as items with scores of 0.09; 0.06; 0.16; 0.22; 0.20, which although not many in number, show that some items are still at extreme difficulty levels. The implication is that this item composition tends to be overly dominated by easy items, while medium and difficult items are still not proportional compared to the ideal summative test target (±20-30% easy, 40–60% medium, 20-30% difficult). To improve instrument quality, items that are too easy ($p > 0.85–0.90$) should be reduced or revised so as not to reduce discrimination power, unless they are indeed intended to measure basic competencies. Conversely, items that are too difficult ($p < 0.25$) should have their wording improved to be more plausible and appropriate to students' cognitive level. With this difficulty distribution balancing, the test is expected to measure participants' abilities more fairly, have good discrimination power, and higher overall reliability.

3.1.5. Distractor Effectiveness

Distractor function analysis on the 35-item tryout (options A-E) shows that out of a total of 140 distractor slots, there are 92 functioning (≈65.7%) and 48 non-functioning (≈34.3%), with an average of 2.63 functioning distractors per item (exceeding the practical target of ≥2), so distractor quality is generally quite good; however, priority repairs are needed for items with <2 functioning distractors, namely 1, 4, 5, 18, 21, 22, 30 for example B1 (key C) and B4 (key E) where three distractors are non-functioning although positively there are no items with 0 functioning distractors; position patterns also show E is rarely the key (A=9, B=7, C=7, D=8, E=4) and the weakest distractor performance is on option B (≈53.6%) and the best on D (≈74.1%) followed by E (≈71.0%), indicating distractor B is relatively less "attractive"; pedagogically, non-functioning distractors are usually too easy to eliminate, not plausible, give clues to the key, or are not appropriate to the cognitive level, so improvements need to emphasize plausibility based on common misconceptions, balance of option form/length, avoidance of form clues, consistency of domain and cognitive level, and more even rotation of key positions.

3.2. Discussion

Item validity analysis using item-total correlation (r_{pB}) provides an overview of the extent to which an item is consistent with the overall test. Items with r_{pB} values ≥ 0.30 are considered valid because they can represent the measured construct well. Conversely, items with $r_{pB} < 0.20$ or negative values are considered invalid, because they most likely do not align with the indicators to be measured and can even reduce overall test reliability. Findings from the data show that in the initial indicator block, many items fall into the invalid category, including items with negative r_{pB}. This indicates weaknesses in item wording, answer keys, or content relevance to the construct. However, a number of other items appear valid with medium to high r_{pB} values. This condition confirms that the quality of items in one test can vary, so the process of item revision and selection must be done.

Theoretically, item-total correlation is a classical approach in Classical Test Theory (CTT) used to determine how representative an item is of the total score (Crocker & Algina, 1986). Haladyna & Rodriguez (2013) emphasize that low or negative correlations usually arise due to incorrect answer keys, ambiguous item stimuli, or distractor options that do not function properly. In addition, the relationship between validity and difficulty level (p) and discrimination power (D) is also important: items that are too easy or too difficult tend to produce low correlations, while items with medium difficulty levels usually have the best discrimination power. Thus, the practical implications of this analysis result are: (1) retaining

valid items, (2) revising items with marginal correlation (0.20-0.29) through improvements in wording and answer options, and (3) replacing or discarding items with low/negative rpB that also have poor discrimination power. This step will improve internal consistency and measurement accuracy of the test instrument.

Based on trial results showing a Cronbach's α reliability coefficient = 0.760, the instrument can be categorized as reliable at a "fair-good" level. Substantively, this value indicates that items in the test have adequate internal consistency in measuring the same construct; in other words, variance in participants' scores is more explained by the target construct than by random measurement errors. In the context of scale development, the range α ≈ 0.70-0.80 is commonly considered adequate for the initial trial phase and usually increases after item cleaning and difficulty level composition restructuring (DeVellis, 2016).

Methodologically, Cronbach's α is influenced by two main things: (1) number of items and (2) average inter-item correlation. Longer scales and/or those with higher inter-item correlations tend to produce larger α. However, pursuing α that is "too high" (>0.90) can actually indicate content redundancy (very similar items), thus reducing construct coverage (Tavakol & Dennick, 2011). In addition, α assumes equivalent item contributions to latent scores; when this assumption is not met, α can be biased, so it is recommended to complement evaluation with alternative coefficients or approaches.

Item discrimination power analysis shows that the majority of items are in the fair category (D = 0.20-0.39), while a small portion falls into the poor category (<0.20) and some are classified as good (≥0.40). This finding shows that the instrument is not yet fully optimal in distinguishing high and low ability participants. Therefore, items with D ≥0.40 should be retained, items with D in the fair category need to be revised (for example, improving stem clarity, reducing overly prominent keys, and strengthening distractors), while items with D <0.20 should be dropped or replaced, especially if also proven invalid.

Theoretically, discrimination power indicates the extent to which an item can distinguish between high and low ability participants. In Classical Test Theory (CTT), one commonly used indicator is the point-biserial correlation (RIT) between item score and total score. As noted by Arifin (2017), items with high discrimination power make important contributions to instrument reliability because they function optimally in detecting variations in participant abilities. The categorization guidelines for D values often used in educational evaluation environments are <0.20 (poor), 0.20-0.39 (fair), and ≥0.40 (good) (Anastasi & Urbina, 1997). Negative values are even considered indications of serious problems, such as key errors, distractors more attractive than correct answers, or stems that create ambiguity.

The results of this analysis imply that instrument development needs to pay attention to distractor quality and difficulty level balance so that discrimination power increases. Previous studies confirm that improvements to items with medium or low discrimination power often successfully increase overall test reliability (Haladyna & Rodriguez, 2013; Tavakol & Dennick, 2011) Thus, recommendations that can be given are to retain items that are already good, revise items that are fair, and replace items that are poor, so that the quality of evaluation instruments can be improved according to educational assessment standards.

Difficulty level (p) analysis is an important step in evaluating test item quality. Classical standards group items into three categories: easy (p > 0.70), medium (0.30-0.70), and difficult (p < 0.30) (Anastasi & Urbina, 1997). Analysis results show that most items fall into the easy category, especially in the first indicator. This indicates that many participants can answer correctly, so these items tend to be less able to distinguish different ability levels of participants. On the other hand, there are also a number of items in the medium category that are quite proportional, as well as several items classified as difficult. Thus, the difficulty level

distribution from this data is not fully balanced, because the dominance of easy items can reduce the discriminative function of the test. If the majority of items are too easy, test results potentially experience a ceiling effect, making them ineffective in distinguishing participants with high and low abilities.

Distractors are wrong answer choices but designed to appear reasonable so they can attract test participants with certain ability levels. Theoretically, distractor quality can be measured through the concept of Distractor Efficiency (DE) which refers to the extent to which distractors function as intended. Distractors are considered functioning if chosen by at least 5% of participants and have a negative correlation with the item total score, while distractors that are rarely chosen or positively correlate with the key are called Non-Functioning Distractors (NFD) (Tarrant et al., 2009). Good distractors are those that can attract more test participants. This distractor efficiency directly affects item difficulty index and discrimination power, making distractor analysis an important part of item quality evaluation (Sharma, 2021).

In modern approaches, distractor analysis is also done with Item Response Theory (IRT) using Option Characteristic Curves (OCC). These curves depict the probability of participants with different ability levels choosing each option. Good distractors are usually chosen more by participants with low abilities and increasingly rarely chosen as ability increases (DeMars, 2010). In addition, development of Differential Distractor Functioning (DDF) methods allows examiners to detect bias at the option level, so distractors are assessed not only from their effectiveness but also from their fairness aspect. Thus, distractors are not just complementary components in multiple-choice questions but are key factors determining the validity, reliability, and fairness of a test. Structured distractor analysis can help test developers improve items to be of higher quality, efficient, and able to measure participant abilities fairly.

4. Conclusion

The tryout instrument has shown adequate internal consistency, but item quality still varies: some items are already aligned with the construct and distinguish participant abilities well, while several others are invalid, have weak discrimination power, and tend to be too easy, thus reducing the discriminative power of the test and indicating problems with wording, answer keys, and distractors. Therefore, good items should be retained; marginal items should be revised by clarifying stems, reviewing keys, and composing more plausible distractors (if necessary reducing the number of options so that those remaining truly function); items that are consistently problematic should be replaced or deleted; difficulty level composition should be restructured to be balanced; content should be realigned with construct indicators; and evaluation should be complemented with alternative model-based analysis. After improvements, conduct retesting on comparable samples and iterate item selection until instrument reliability, validity, and discriminative function improve according to assessment standards.

Based on these findings, several actions are recommended, first, effective items should be retained while marginal items require revision through clarifying question stems, reviewing answer keys, and developing more plausible distractors which potentially reducing option numbers to ensure functional alternatives; consistently problematic items should be replaced or eliminated; the difficulty level distribution should be rebalanced; content alignment with construct indicators must be strengthened; and classical analysis should be supplemented with model-based approaches, followed by retesting on comparable samples until reliability, validity, and discrimination meet assessment standards. Future research should employ

complementary methods such as Item Response Theory (IRT) for deeper insight into item functioning, conduct qualitative investigations like think-aloud protocols to validate the PCK construct being measured, and explore the application and refinement of this PCK-based assessment model across various vocational disciplines.

5. References

Adipat, S., Chotikapanich, R., Laksana, K., Busayanon, K., Piatanom, P., Ausawasowan, A., & Elbasouni, I. (2023). Technological Pedagogical Content Knowledge for Professional Teacher Development. *Academic Journal of Interdisciplinary Studies*, *12*(1), 173. https://doi.org/10.36941/ajis-2023-0015

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Prentice Hall.

Anggreini, D., & Darmawan, C. A. (2017). Analisis Kualitas Soal Try Out Ujian Nasional Dengan Menggunakan Aplikasi Program Anates. *JP2M (Jurnal Pendidikan Dan Pembelajaran Matematika)*, *2*(1), 20. https://doi.org/10.29100/jp2m.v2i1.213

Arifin, Z. (2017). *Evaluasi Pembelajaran: Prinsip, Teknik dan Prosedur*. Remaja Rosdakarya.

Crocker, L. M., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart, and Winston. https://books.google.co.id/books?id=tfgkAQAAMAAJ

DeMars, C. (2010). *Item Response Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

DeVellis, R. F. (2016). *Scale Development: Theory and Applications*. SAGE Publications.

Dewi, M. S., Setyosari, P., Kuswandi, D., & Ulfa, S. (2020). Analysis of Kindergarten Teachers on Pedagogical Content Knowledge. *European Journal of Educational Research*, *volume-9-2*(volume-9-issue-4-october-2020), 1701–1721. https://doi.org/10.12973/eu-jer.9.4.1701

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

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Sharma, L. R. (2021). Analysis of Difficulty Index, Discrimination Index and Distractor Efficiency of Multiple Choice Questions of Speech Sounds of English. *International Research Journal of MMC*, *2*(1), 15–28. https://doi.org/10.3126/irjmmc.v2i1.35126

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(2), 4. https://doi.org/10.2307/1175860

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, *9*(1), 40. https://doi.org/10.1186/1472-6920-9-40

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd