Jurnal
Kiprah Pendidikan
Volume 4 Nomor 4 Oktober 2025, Pg. 715-724

ISSN 2810-0443 (print) | 2827-8909 (online)
https://kpd.ejournal.unri.ac.id
https://doi.org/10.33578/kpd.v4i4.p715-724

DEVELOPING A PEDAGOGICAL CONTENT KNOWLEDGE (PCK) ASSESSMENT FOR PRE-SERVICE ARABIC LANGUAGE TEACHERS

PENGEMBANGAN TES *PEDAGOGICAL CONTENT KNOWLEDGE* (PCK) UNTUK MAHASISWA PENDIDIKAN PROFESI GURU BIDANG STUDI BAHASA ARAB

Laily Maziyah[1], Muhammad Alfan[2], Nur Faizin[3]
[1,2,3] Universitas Negeri Malang, Malang, Indonesia
E-mail: laily.maziyah.fs@um.ac.id [1], muhammad.alfan.fs@um.ac.id[2], nur.faizin.fs@um.ac.id [3]

Submitted
19 September 2025

Accepted
25 September 2025

Revised
12 Oktober 2025

Published
30 Oktober 2025

Kata Kunci:
Analisis butir;
Validitas;
Reliabilitas;
Daya beda;
Pengecoh

Keyword:
Item analysis:
Validity;
Reliability;
Discrimination index;
Distractor

Abstrak

Penelitian ini bertujuan untuk mengembangkan tes PCK mahasiswa peserta Pendidikan Profesi Guru (PPG) bidang studi Bahasa Arab. Penelitian ini menganalisis kualitas butir soal try out Bahasa Arab dengan menggunakan parameter tingkat kesulitan, daya beda, validitas, fungsi distraktor, serta reliabilitas instrumen. Data diperoleh dari hasil pengerjaan 291 peserta tryout dengan jumlah soal 35 butir pilihan ganda. Analisis dilakukan melalui perhitungan indeks kesukaran, daya beda, korelasi butir-total, evaluasi distraktor, dan uji reliabilitas Cronbach's Alpha. Hasil penelitian menunjukkan bahwa sebagian besar butir berada pada kategori sedang, dengan reliabilitas instrumen yang cukup baik. Selain itu, ditemukan sejumlah soal yang perlu direvisi untuk meningkatkan kualitas instrumen evaluasi.

Abstract

This study aims to develop an PCK assessment for pre-service Arabic language teachers. The analysis focused on the quality of the Indonesian-language try-out items, considering parameters such as difficulty level, discrimination power, validity, distractor effectiveness, and instrument reliability. Data were collected from the responses of 291 try-out participants to 35 multiple-choice items. The analysis involved calculating difficulty indices, discrimination indices, item-total correlations, evaluating distractor effectiveness, and testing reliability using Cronbach's Alpha. The findings revealed that most items fell into the moderate difficulty category, while the instrument demonstrated a fairly good level of reliability. Nonetheless, several items were identified as requiring revision to enhance the overall quality of the evaluation instrument.

Citation :
Maziyah, Lily, dkk. (2025). Developing a Pedagogical Content Knowledge (Pck) Assessment for Pre-Service Arabic Language Teachers. *Jurnal Kiprah Pendidikan, 4* (4), 715-724. DOI: **https://doi.org/10.33578/kpd.v4i4.p715-724**.

INTRODUCTION

Teacher Professional Education (Pendidikan Profesi Guru/PPG) is one of the strategic programs in an effort to improve the quality of education in Indonesia. As stipulated in Law Number 14 of 2005 concerning Teachers and Lecturers, the duties of teachers include planning and implementing learning, evaluating, and assessing and evaluating students. Therefore, competency assessment and evaluation are critical elements in ensuring that PPG graduates have pedagogical, professional, social, and personality abilities that meet standards.

Pedagogical Content Knowledge (PCK) has become a major focus in the study of teacher education and professional development, given its long-recognized role as a crucial indicator in determining the quality of an educator (Nilsson & Loughran, 2012; Park & Suh, 2015). PCK

ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724
Jurnal Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724

represents the essence of effective learning practices, as it reflects a distinctive integration between content and pedagogical aspects. This integration forms the foundation for teachers' professional understanding of how to structure, sort, and represent a particular topic, issue, or question in a systematic and meaningful way in the context of learning (Shulman, 1987). Along with the shift in the educational paradigm towards Outcome-Based Education (OBE) or learning outcome-oriented education, the relevance of assessment questions is increasing. Badriah & Robandi (2023) explained that assessments do not just measure what has been taught, but must be linear with learning objectives, the achievement of competencies of program graduates, and must be able to provide useful feedback for both students and PPG institutions..

In reality, in the field, challenges are often found in assessment instruments in the form of questions that are not valid and reliable, poor distractions, poor differentiation between high-ability and low-ability students, or ambiguous question language. Without good instruments, assessment results may not reflect students' actual competencies, which has a negative impact on educational decision-making and learning improvement.

On the other hand, assessment technology and information systems have begun to be used as part of the OBE implementation strategy to assist study programs in monitoring graduate learning outcomes, making continuous improvements, and accreditation. This shows that it is not only the quality of the questions themselves that matters, but also how the assessment questions are managed, measured, and interpreted in a larger system.

The development of a Pedagogical Content Knowledge (PCK) test for prospective teachers or educators is very important for several main reasons. First, the importance of PCK for Effective Teaching. PCK is a category of teacher knowledge that includes how teachers successfully deliver subject matter to students in classroom practice. It is considered an important predictor of teacher quality and a critical element for effective preparation. Teachers with strong PCK are known to be successful in teaching certain topics. PCK also helps teachers transform their content knowledge into a pedagogically strong form, yet adaptive to students' knowledge, comprehension levels, and learning difficulties.

Second, improving student learning outcomes. Research shows that PCK can improve students' academic performance through better quality instruction. The impact of PCK on students' academic achievement has proven to be stronger than Content Knowledge (CK) alone. Therefore, assessing and developing PCK is essential to ensure teachers can design cognitively challenging learning experiences.

Third, identify shortcomings and gaps in the knowledge of prospective teachers. PCK tests such as Content Representation (CoRe) can reveal significant deficiencies in participants' PCK, especially in understanding students' difficulties and planning effective teaching methodologies. This instrument allows teacher education programs to identify areas where prospective teachers need further support.

Fourth, supporting Professional Development. PCK develops during the professionalization process, and tests can help measure the extent to which teachers have developed their PCK and how they can improve it. These tools can be integrated into teacher education programs to address gaps and better prepare prospective teachers to teach complex mathematical concepts, such as fractions as operators.Fifth, informing the curriculum design of teacher education programs. Understanding the type of knowledge that prospective teachers possess is a prerequisite that can help improve the curriculum design of educational degree programs. The results of the PCK assessment can provide recommendations for the development of teacher training programs that focus on these important characteristics.

Jurnal
Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724
ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724

Overall, the development of a valid and reliable PCK test ensures that teacher education programs can effectively foster and evaluate the ability of prospective educators to teach effectively, ultimately contributing to improved student learning outcomes. Thus, the development of assessment questions for PPG students is not only relevant, but urgent, in order to produce valid, reliable instruments, able to distinguish student performance, support the achievement of competencies targeted by PPG, facilitate feedback for lecturers and students, and can be applied in the context of OBE and the broader quality assurance system.

The development of test instruments in this study is carried out through a series of systematic procedures that include the collection, analysis, and interpretation of test results data as a basis for decision-making and quality assurance of assessment instruments. This process has direct implications for the readiness of participants to face the national exam. To ensure that the inferences resulting from test scores are valid and reliable, the instrument used must provide empirical evidence of validity the conformity between the score and the construct being measured and reliability the internal consistency between items. In addition, grain characteristics such as the level of difficulty, differentiating power, and effectiveness of the divertor must also meet adequate psychometric standards (Azwar, 2020).

Departing from this framework, this study focuses on the analysis of Arabic tryout questions, which function both as a simulation before the exam and as a diagnostic tool to identify the strengths and weaknesses of the participants. The implementation of experiments on instruments is a crucial stage because it provides opportunities for researchers and policy makers to evaluate the quality of question items, examine the effectiveness of the distractors, and estimate the validity and reliability of the instruments before they are widely applied (Arikunto, 2018; Sudjana, 2019).

METHOD

This study uses a quantitative approach with question item analysis techniques based on classical test theory. The subjects of the study were 291 Arabic tryout participants who worked on 35 multiple-choice questions. The data was analyzed through stages: (1) calculating the difficulty index (p); (2) calculating the differential power index (D); (3) assess the validity of the item with the item-total correlation corrected; (4) evaluate the distractor through the distribution of students' answer choices; and (5) test the overall reliability of the test using the Alpha Cronbach coefficient (Azwar, 2020; Arikunto, 2018).

Trial Design and Procedures. The trial was carried out on 291 participants with a 35-point multiple-choice instrument (5 options, 1 correct key). Each correct answer is given a score of 1 and false 0. The analysis was performed to assess the feasibility of the grain, the function of the distractor, and the internal reliability.

Rules for Item Eligibility Decisions. (a) Used without revision: valid (≥ 0.30), $D \geq 0.40$, functional distractor; (b) Used with revision: valid but D 0.20–0.39 and/or weak distractor; (c) Dropped: invalid (< 0.30) and/or $D < 0.20$ (Azwar, 2020).

RESULTS AND DISCUSSION

Difficulty Level of Question Item

Analysis of the difficulty level of the question items shows that there is a quite striking variation in the category. Some questions are easy because they only test the understanding of basic concepts, while others are difficult because they require students to apply the concepts to more complex

ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724
Jurnal Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724

situations. This variation suggests that the questions are well-designed to measure students' level of comprehension as a whole, so as to provide an accurate picture of the student's abilities in the material being tested (Febilia et al., 2024). Some items are in the easy category with a high proportion of correct answers ($p > 0.80$), for example in items 1 and 2. This condition indicates that some questions tend to be less challenging for participants, so they are not optimal in mapping the variety of PPG students' abilities. However, there are also a number of items that are in the medium category ($0.30 \leq p \leq 0.80$), such as items 3 and 4, which are more effective in distinguishing the participants' abilities. It is this medium category item that, according to classical theory, contributes the most to the reliability and validity of the instrument.

The relatively dominant composition of easy grains has the potential to weaken the discriminating power of participants (Sri, n.d.). Therefore, it is necessary to review the items that are considered easy so that the test can be more effective in accurately measuring the variation in PPG students' abilities. The ideal instrument should have a higher proportion of questions with moderate difficulty, because this group is the most diagnostic in selectingparticipantsbased onability (Gusmizain, 2022; Wijaya et al., 2024). Therefore, revisions to easy items are needed, either by deepening the material or strengthening the trick options, to be more balanced and in accordance with the test development guidelines. Thus, improving the quality of test instruments can help in identifying participants who have the ability to meet the set standards. In addition, the revision of easy items can also improve the validity and reliability of the test overall.

From a total of 35 questions, the results of the difficulty level analysis showed the distribution as follows:

- Easy category ($p > 0.80$) as many as 11 items (34%), for example items 1, 2, 5, 6, 7, 8, 9, 10, 11, 14, 25, and 26. These items were answered correctly by the majority of participants, making it less challenging to map the variation in PPG students' abilities.
- Medium Category ($0.30 \leq p \leq 0.80$) as many as 22 items (66%), such as items 3, 4, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 27, 28, 29, 30, 31, 32, 33, 34, and 35. This group of questions is more effective for selecting the ability of test takers.
- Difficult category ($p < 0.30$) was found as many as 2 items (6%) on this instrument.

The absence of difficult items and the dominance of easy questions have implications for a decrease in the sensitivity of the instrument to high-ability participants. In fact, medium-difficult category questions are needed to optimize the reliability of the test. Therefore, the composition of the questions still needs to be reviewed to be more balanced. A balanced composition of questions between medium and difficult difficulty levels will help improve the quality of the test and can provide a more accurate picture of the participants' abilities (Jesika & Andika, 2024). Thus, a review of the composition of the questions is an important step in the development of effective evaluation instruments.

Power of Difference Item Question

The results of the differential power analysis showed various qualities. Some items have sufficient to good discriminating power, for example items 3 and 4 which are classified as being able to separate high-ability and low-ability students relatively effectively. However, there are also a number of items with low differentiation, even close to the weak category ($D < 0.20$). Items with low differentiation basically do not provide meaningful information related to the participants' abilities, because the questions are answered correctly or incorrectly almost equally by all groups.

The dominance of grains with sufficient differentiation indicates that this instrument has the

Jurnal
Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724
ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724

potential to be used, but the existence of grains with low discrimination must be addressed immediately. According to Surbakti (2025), substantial revision or replacement of new items is an important step to increase reliability (Surbakti, 2025). If a test has a number of items with low differentiation that are mostly answered correctly by all participants, then the test results will not provide accurate information regarding individual abilities (Kurniawan & Pradipta, 2023). On the other hand, if there are items with high differentiation but only a few correct answers by participants, this can also result in inconsistent or invalid information (Mulvia et al., 2021). A stronger ditractor, editorial clarity, and suitability of material indicators can be a solution so that the differentiation power increases and the function of the instrument as a selection tool becomes more optimal (Ida & Musyarofah, 2021).

The differential power analysis (D) revealed the quality of the items in distinguishing high and low ability participants. The results of the distribution are as follows:

- Good–Very Good Category ($D \geq 0.40$) as many as 13 items (37%), including items 3, 12, 13, 15, 16, 18, 19, 22, 23, 28, 29, 33, and 34. These questions effectively separated the high-ability participants from the low-ability participants.
- Sufficient Category ($0.20 \leq D < 0.40$) as many as 18 items (51%), for example items 1, 2, 4, 5, 6, 7, 10, 17, 20, 21, 24, 27, 30, 31, 32, 35, and others. These items can still be used, but they need to be revised to be more optimal.
- Weak category ($D < 0.20$) as many as 4 items (12%), namely items 8, 9, 11, and 25. Items with low discriminating power should be repaired or replaced.

Potential to reduce the reliability of the instrument. Therefore, it is necessary to revise the items with the Weak category in order to improve the overall reliability of the instrument. Nevertheless, the results of this analysis provide a fairly positive picture of the overall quality of the questions. By making improvements to items that have low discrimination, it is hoped that this evaluation instrument can provide more accurate and reliable results in measuring student competence.

Validity of Question Items

The validity aspect also shows an interesting pattern. Most of the items show an adequate item–total correlation ($r\_it \geq 0.30$), so it can be said to be quite valid in measuring the targeted construct. These items can be retained without substantial revisions. However, there are also a number of invalid items ($r\_it < 0.30$), such as items 1 and 5, that do not consistently contribute to the total score. This low validity can occur due to ambiguous question redaction, extreme difficulty levels, or answer options that do not function as they should (Febilia et al., 2024). With the findings on the level of validity of the items in this measurement instrument, it is necessary to take regular improvement and update actions so that the quality of the measuring instrument is maintained and relevant to the intended construction. This aims to ensure that the data obtained from the measurement instrument is reliable. So that decisions made based on this data can provide accurate and appropriate results.

This condition is in line with the principle affirmed by international assessment standards, that invalid items must be re-evaluated or replaced so that the quality of the instrument is maintained (Septi, 2018). This effort is important, not only statistical technical, but also pedagogical reflection so that test items really measure abilities that are relevant to the academic potential of prospective teachers. This shows how important the validity of measurement instruments is in the world of education, especially in the selection process of prospective teachers. By ensuring that the test items

ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724
Jurnal Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724

truly reflect the relevant abilities, it can be ensured that the teacher selection process can be carried out more effectively and efficiently. In addition, the validity of the measurement instrument will also ensure that the selected teacher candidates really have academic potential in accordance with their educational needs.

- Valid items ($r\_it \geq 0.30$) as many as 20 items (57%), including items number 2, 3, 7, 12, 13, 15, 16, 17, 18, 19, 20, 21, 22, 23, 28, 29, 32, 33, 34, and 35. These items can be retained as the core part of the instrument.

- Invalid items ($r\_it < 0.30$) as many as 15 items (43%), such as items 1, 4, 5, 6, 8, 9, 10, 11, 14, 24, 25, 26, 27, 30, and 31. This low validity can be due to ambiguous redaction, extreme difficulty levels, or distractor options that don't work optimally.

The high proportion of invalid items (43%) indicates the need for substantial revisions so that the instrument is more accurate and consistent in measuring students' academic potential. Although there are invalid items, the proportion of valid items is still high enough to maintain the instrument as a core part. Substantial revision may not be necessary if only 43% of the item is invalid.

The Results of the Analysis of the Distractors

Overall, from the 35 questions analyzed, the results can be summarized as follows:

- Difficulty level: 12 easy (34%), 23 moderate (66%), 0 difficult (0%). Most items were moderately difficult, and none were considered difficult. This test was well-designed to fairly and accurately measure prospective Arabic language teachers' knowledge and skills. This test can provide accurate predictions of prospective Arabic language teachers' academic success.

- Power differences: 13 good grains (37%), 18 good grains (51%), 4 weak grains (12%). Most of the question items have a good or sufficient difference. Although there are several question items that have weak differentiation, the proportion of question items with good and sufficient differentiation is still higher. This shows that this test is able to distinguish between students who have different knowledge and skills more effectively. Thus, this test is expected to provide more accurate information in evaluating the abilities of prospective Arabic teachers.

- Validity: 20 valid items (57%), 15 invalid items (43%). The proportion of valid question items is still higher than invalid, further evaluation needs to be carried out to correct invalid question items. The validity of the test is important to ensure that the test actually measures what it is supposed to measure. That way, it is hoped that this test can provide a more accurate picture of the ability of prospective Arabic teachers.

This instrument has the power to dominate medium grains and most of the grains are valid with good differentiation. However, the weakness is still seen in the number of easy questions, the number of invalid items, and the existence of items with low discriminating power. To improve quality, the strategies that can be pursued are (Solichin, 2017):

1. Maintain valid and high-potency items.
2. Revise the grain enough to go up for good, especially by strengthening the distractor.
3. Replace invalid items and weak differentiation with new questions based on more appropriate indicators.

With this improvement step, the instrument is expected to be more reliable, valid, and able to select the potential academic success of PPG Arabic students more accurately. In addition, it is also

Jurnal
Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724
ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724

important to hold training for teachers in compiling quality exam questions. Thus, they can understand more deeply about the importance of paying attention to the validity and difference in each question item that is prepared. In addition, the exam supervisor also needs to be involved to ensure that the implementation of the strategy that has been set can run well and effectively. Thus, the exams held can really be a reliable evaluation tool in measuring the ability of PPG Arabic students.

Question Category Analysis Analisis

The following is presented the categorization of questions based on the number, percentage and number of questions.

Table 1. Question categories

Question Categories	Total	Percentage	Question number
Valid & used without revision	6	17,10%	2, 3, 12, 13, 25, 26
Invalid but still usable (revision)	12	34,30%	4, 6, 10, 14, 15, 16, 18, 20, 22, 23, 27, 32
Invalid & must be dropped	17	48,60%	1, 5, 7, 8, 9, 11, 17, 19, 21, 24, 28, 29, 30, 31, 33, 34, 35
Total	35	100%	

1. Valid Items & Used Without Revision

A total of 6 items (17.1%) showed validity ≥0.30, "sufficient" differentiation (≥0.35), and difficulty level in the medium category. This condition is ideal because the question is able to distinguish between high-ability and low-ability students. These items can be retained without any significant changes.

2. Invalid Items but Can Still Be Used with Revision

There were 12 items (34.3%) whose validity was <0.30 but still had a "sufficient" difference (≥0.30) with the level of difficulty in the medium category. This means that potentially the grain is still useful, but the diverter needs to be improved to make it more functional. The revision is focused on the trickster option that the student did not choose or is too clearly wrong, so that the differentiation can increase.

3. Invalid Items & Dropped

A total of 17 items (48.6%) had to be dropped. Most of these items have a validity score of <0.20 with a "bad" differentiation (<0.25). There are also problems with extreme difficulty levels (too easy or too difficult), which results in the instrument not being able to distinguish the participants' abilities. This item should not be used again because the repair will require a complete reconstruction, not just a trickster revision. By eliminating invalid and meaningless items, the differentiation of the evaluation instrument can be increased and provide more useful information in measuring participants' abilities.

ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724
Jurnal Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724

Linkage to Reliability

Previous results showed that the reliability of the instrument (Cronbach's Alpha = 0.78) was in the good category but not optimal. The high number of invalid questions (48.6%) explains why reliability does not reach ≥0.80. Therefore, it is necessary to revise the measurement instruments used. The revision was carried out by streamlining the number of invalid questions. This is done so that the reliability of the measurement instrument can reach the desired level. If the 17 dropped items are removed, reliability is expected to increase. If the 12 items that can still be revised are repaired the tractor, reliability can increase more significantly, the possibility of reaching the very good category (≥0.80). However, increased reliability can also be influenced by other factors such as internal and valid consistency.

Quality Distribution of Distractors

The distribution of distractor quality in the 35 questions above can be presented in the following table:

Table 2. Quality Distribution of Distractors

Question Categories	Question Number	Total	Percentage
Works well (no revision needed)	9, 10, 21, 22, 26	5	14,30%
Partially functional (revision needed)	1, 2, 3, 4, 5, 6, 8, 12, 13, 14, 15, 16, 18, 23, 24, 25, 28, 29, 30, 31, 32, 33, 34, 35	24	68,60%
Not working (very bad/dropped)	7, 11, 17, 19, 20, 27	6	17,10%
Total	—	35	100%

Data Interpretation

The distraction functions well (14%) → means that the deceitful has worked according to the principle (chosen ≥5–10% of low-ability participants and rarely high-ability participants selected). This item supports the validity of the construct. This item supports the validity of the construct. And provide useful information to improve the quality of the test. And provide useful information to improve the quality of tests and ensure that they can measure accurately.

Partially functional distractions (69%) → deceitful still attract some participants, but the distribution is not ideal. For example, there are options that are too weak or rarely chosen. This item can still be used, but it needs to be revised on certain options. So that the test can produce more valid and reliable data. So that the test can produce more valid and reliable data for participants who take the exam. So that the test can produce more valid and reliable data for participants who take the exam. This is important.

The distraction did not work (17%) → the trickster was not chosen at all or instead chosen by highly capable participants. This condition is dangerous because it can reduce the power of difference and validity. This category should be completely revised or dropped. This category should be completely revised or dropped in order to maintain fairness in the measurement of participants' abilities. in the measurement of participants' abilities. Maintaining fairness is essential to ensure accurate outcomes.

Relevance to Validity & Differentiation

Items with good distractors generally also show validity of ≥0.30 and differentiation of ≥0.40.

Jurnal
Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724
ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724

can demonstrate that such measurement instruments are reliable to provide consistent and objective results. Thus, researchers can trust that the data obtained from these instruments can be relied upon to ensure the accuracy of the research results. Therefore, the validity of the measurement instrument can be considered as one of the key factors in ensuring reliability.

Items with a partially functional distractor; Usually the validity is sufficient (≥0.30), but the differential power is low (0.20–0.39). A change in the power of the deception can increase the power of difference. However, it can also improve the overall validity of the measurement instrument. However, it can also improve the overall validity of the measurement instrument, especially if the measurement revision is done carefully and takes into account various relevant aspects.

Items with non-functioning diverters often have a validity of <0.30 and a differentiation power of <0.20, making them more appropriate to abort or rearrange. These items can affect the validity and reliability of the measurement instrument as a whole. These items can affect the validity and reliability of the measurement instrument as a whole, so it is necessary to conduct a thorough review of these items.

Integration with Reliability

When compared withreliability data, the previous Cronbach Alpha coefficient was 0.78. This distribution of distractor quality explains why reliability has not yet reached the "excellent" category (≥0.80). This suggests that improvements need to be made to the quality of the diverter to achieve the desired level of reliability. Improvements to the quality of the diverter can be made through the revision or replacement of the poor diverter. Here are the steps that can be taken to improve the quality of the dystractor in the test. (1) Deleting 6 items does not work → reliability can be improved. (2) Revising 24 grains with a weak divertor will strengthen the internal consistency of the instrument.

CONCLUSION AND RECOMMENDATION

If viewed as a whole, the test instrument for the potential academic success of PPG Arabic students already has a good foundation, with a number of valid and adequate points. However, there are still weaknesses in the form of easy item dominance, the existence of items with low discriminating power, and some invalid questions. This condition indicates that the instrument requires further review before full use.

Strategic steps that can be taken are (1) maintaining valid items with high differentiation, (2) revising items with sufficient categories to move up to the good category, and (3) replacing invalid items with new questions that are prepared based on sharper competency indicators. Thus, this test is expected to be able to be a fair, reliable, and valid selection instrument in predicting the academic success of prospective Arabic teacher students.

Based on the results of the analysis of question items, the quality of the instrument can be categorized into three main groups. First, there are 5 question items (14%) that can be used without revision. These questions are considered valid, have good differentiation, and are supported by a dynamic that functions optimally. This shows that the item has met the criteria of a good instrument, so that it is able to effectively distinguish students with different levels of mastery of the material.

ISSN 2810-0443 (print) | 2827-8909 (online)
https://doi.org/10.33578/kpd.v4i4.p715-724
Jurnal Kiprah Pendidikan
Volume 4 Nomor 4 October 2025, Pg. 715-724

BIBLIOGRAPHY

Arikunto, S. (2018). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.

Azwar, S. (2020). *Reliabilitas dan validitas*. Yogyakarta: Pustaka Pelajar.

Badriah, & Robandi, B. (2023). Outcome-based education pada kurikulum merdeka: Linearitas pembelajaran dengan asesmen untuk mencapai tujuan pembelajaran. *Paedagoria: Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, 14(2). https://journal.ummat.ac.id/index.php/paedagoria/article/view/16042

Febilia, C., Cahyani, R., Luqman, A., & Vivi, N. (2024). Analisis butir soal pilihan ganda pada elemen akuntansi keuangan guna mengoptimasi evaluasi menggunakan Anates V4.0. *Jurnal Nasional Sains dan Inovasi (JNSI)*, 1(1). https://ejournal.nlc-education.or.id/index.php/JNSI/article/view/95

Gusmizain, A. (2022). Karakteristik Butir Soal Tes Mata Kuliah Matriks Ruang Vektor Mahasiswa Matematika. Jurnal Evaluasi Pendidikan, 13(2), 127–130. https://doi.org/10.21009/jep.v13i2.28328

Ida, F. F., & Musyarofah, A. (2021). Validitas dan Reliabilitas dalam Analisis Butir Soal. *AL-MU'ARRIB: JOURNAL OF ARABIC EDUCATION*, *1*(1), 34–44. https://doi.org/10.32923/al-muarrib.v1i1.2100

Jesika, & Andika. (2024). Penerapan implementasi soal tes pilihan ganda di SDN Macajah 2. *Jurnal Inovasi Pendidikan Dasar (JIPDAS)*, 4(1). https://ejournal.lpipb.com/backup_ejournal_v1/index.php/jipdas/article/view/497

Kurniawan, R., & Pradipta, A. W. (2023). Tingkat Kesulitan dan Daya Beda Butir Soal Ujian Akhir Semester Mataku liah Penelitian Pendidikan. Jurnal Pendidikan, 11(2), 234–341. https://doi.org/10.36232/pendidikan.v11i2.3999

Mulvia, R., Ramalis, T. R., & Efendi, R. (2021). Mendeteksi Keajegan Butir Tes Dengan Fungsi Informasi. Jurnal Pendidikan Indonesia, 2(1), 72–84. https://doi.org/10.36418/japendi.v2i1.66

Nilsson, P., & Loughran, J. (2012). Exploring the development of pre-service science elementary teachers' pedagogical content knowledge. *Journal of Science Teacher Education, 23*(7), 699–721. https://doi.org/10.1007/s10972-011-9239-y

Park, S., & Suh, J. (2015). From portraying toward assessing PCK: Drivers, dilemmas, and directions for future research. In A. Berry, P. Friedrichsen, & J. Loughran (Eds.), *Re-examining pedagogical content knowledge in science education* (pp. 104–119). Routledge.

Septi. (2018). *Buku ajar mata kuliah asesmen pembelajaran*. Umsida Press. https://press.umsida.ac.id/index.php/umsidapress/article/download/978-602-5914-21-8/822

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Solichin. (2017). Analisis daya beda soal, taraf kesukaran, validitas butir tes, interpretasi hasil tes dan validitas ramalan dalam evaluasi pendidikan. *Jurnal Evaluasi Pendidikan*, 8(1). http://download.garuda.kemdikbud.go.id/article.php?article=563543&val=9596

Sri, I. (2023). *Dasar-dasar evaluasi pembelajaran* (Tesis, UIN Raden Intan Lampung). UIN Repository. https://share.google/qfv59hUnQpbinoema

Surbakti. (2025). Analisis kualitas butir soal pada uji coba evaluasi pembelajaran matematika. *Cendekia: Jurnal Pendidikan dan Pembelajaran*, 5(1), 15–28. https://jurnalp4i.com/index.php/cendekia/article/view/4097

Wijaya, A., Ismady, M. W., & Rosdiana. (2024). Analisis Pengukuran Kompetensi Pendidikan Profesi Guru Dalam Jabatan. Indonesian Journal of Islamic and Social Science, 2(2), 67–79. https://doi.org/10.71025/hepjer05