Cakrawala Pendidikan
Jurnal Ilmiah Pendidikan
Vol. 43 No. 1, February 2024, pp. 192-203
https://journal.uny.ac.id/index.php/cp/issue/view/2613
DOI: https://doi.org/10.21831/cp.v43i1.53558

The quality of teacher-made summative tests for Islamic education subject teachers in Palembang, Indonesia

Evy Ratna Kartika Wati[1]*, Yanti Karmila Nengsih[1], Ciptro Handrianto[2], M. Arinal Rahman[3]
[1]Universitas Sriwijaya, Indonesia, [2]Sultan Idris Education University, Malaysia,
[3]University of Szeged, Hungary
*Corresponding Author: exyrkwaty@gmail.com

ABSTRACT

Criticism of exams can be used to gauge student achievement for graduation. This study examined the quality of summative tests (ST) created by senior high school teachers in Palembang, Indonesia, specifically focusing on the Islamic Education subject. The evaluation criteria included item validity, reliability, discrimination index, and disruptive effectiveness. The analysis involved 800 answer sheets from 20 teachers. Results indicated that while 20 teachers achieved high reliability, two struggled with poor reliability in terms of disruptive effectiveness. Respondent 11 faced challenges with only 28% valid items and a moderate Cronbach's alpha. Additionally, the disruptive effectiveness and discrimination index were poor. These findings suggest a need for teacher training to enhance skills in crafting and administering high-quality summative tests. The implications of these findings extend to improving teacher training and ensuring the effectiveness of summative assessments in gauging student achievements for graduation. The research contributes valuable insights into the complexities of teacher-created exams and offers a basis for enhancing the overall quality of education assessment practice.

Keywords: quality, summative test, validity, reliability, item difficulty level, item discrimination index, item disruptive impact

Article history

Received:	*Revised:*	*Accepted:*	*Published:*
4 April 2023	*20 October 2023*	*3 January 2024*	*28 February 2024*

Citation (APA Style): Wati, R. R. K., Nengsih, Y. K., Handrianto, C., & Rahman, M. A. (2024). The quality of teacher-made summative tests for Islamic education subject teachers in Palembang Indonesia. *Cakrawala Pendidikan: Jurnal Ilmiah Pendidikan*, *43*(1), 192-203. DOI: https://doi.org/10.21831/cp.v43i1.53558

INTRODUCTION

The quality of teacher-created exams implies that teachers have clear assessment criteria, consistent goals and assignments, and legitimate and competent assessment processes (Kartowagiran et al., 2019; Koswara et al., 2021; Widiana et al., 2021). Teacher evaluation illustrates the teaching profession's quality standards. The research found that the quality of this profession significantly affects teachers' attempts to improve formative and summative assessments (Cobbold & Wright, 2021; Moeti, 2016).

The National Education Committee of the Republic of Indonesia prepares for the national exam or standard test. At the same time, teachers take the summative test to determine primary school, junior high school, and senior high school pass rates. Through training and experience, teachers may improve test-taking talents, skills, tenacity, and originality. Teacher-created examinations are based on the provincial education office's test framework for elementary and high schools. The test framework describes the distribution of test items, including complexity (easy, medium, and complex), cognitive capacity, and test form (Malone, 2010).

First, teachers' examinations and assessments lack assessment and measurement understanding (Mokshein et al., 2019). If inexperienced teachers evaluate, it is not unexpected that exam quality is poor (Blakemore, 2012). Sometimes, unclear test findings lead to erroneous

policies or outcomes. Second, summative tests are poorly prepared and administered. New teachers who lack pedagogical skills and student performance evaluation may replace departing teachers. Experienced teachers create better exam questions. Formative test-takers are likelier to administer excellent summative examinations to gauge student success (Laverty et al., 2012). According to Hagen (2020), student performance evaluation is a method of acquiring and giving meaningful information about students' accomplishments and abilities. Accurate knowledge leads to good outcomes. Educators must gather and share information. Third, test quality influences the accuracy of the student accomplishment assessment (Jusuf et al., 2019; Herlina et al., 2021). The teacher's function as a test-maker is crucial. If teachers can create formative exams properly throughout the learning process, they should not have substantial difficulty creating summative assessments.

Teachers play a pivotal role in enhancing education quality, as Armiati et al. (2020) emphasize. Their study at SD Negeri 023 Semoga Jaya highlights the positive impact of incorporating test-understanding quality questions into the midsemester exam, leading to improved performance indicators with an average score of 75. Ridho (2018) also stresses the importance of evaluation in Arabic summative tests, exploring objectives, principles, and techniques. The article provides comprehensive insights into evaluating Arabic summative tests, covering test and non-test methods. In response to the challenge of low test understanding of the quality of the summative tests in junior high school students, Herawati (2021) examines teacher professionalism in crafting such questions before and after training. Surveying 17 teachers from 11 junior high schools in North Padang District, the study reveals that 70% of teachers, post-training, adeptly designed tests to understand quality questions, indicating an enhanced grasp of the training material. Therefore, the quality questions have shown positive outcomes regarding improved performance indicators, but the investigation is crucial to address challenges and enhance teachers' ability to create high-quality questions.

The inquiry is whether the test prepared by senior high school teachers in Palembang fulfills the test content structure, whether the distribution of test items covers the complete class topic, and whether the test items prepared by teachers are of good quality. No one knows how teachers prepare exam questions, whether they use question banks, reference books, or write them themselves.

All these questions require responses to preserve the summative test's efficacy. Teachers are impacted by test elements while creating excellent assessments (Chan, 2018; Supriyadi et al., 2019). Factors influencing teacher test quality and the restrictions teachers confront in preparing for the exam must also be investigated to enhance the study of teacher test quality. Therefore, research must assess the quality of summative examinations created by teachers in Palembang, Indonesia. Summative test quality information provided by representative teachers is anticipated to answer criticism of exams used to gauge student achievement for graduation in Palembang and across Indonesia. By examining the test's quality, the teacher may self-reflect on whether the summative test met excellent quality or whether they made a decent language test. If a test is valid and reliable, it reflects well on students' competency. The validity, test reliability, item difficulty level, item discrimination index, and disruptive impact of the test items may assist the teacher in recognizing successful items. The study aimed to assess the quality of summative examinations created by teachers in Palembang, Indonesia. The study aims to determine if the test content structure is fulfilled, if the distribution of test items covers the complete class topic, and if the test items provided by teachers are of high quality. The study also investigated the factors influencing teacher test quality and teachers' limitations in preparing for the exam. By examining the quality of the summative test, the study hopes to provide information that can address criticisms of exams used to gauge student achievement for graduation in Palembang and across Indonesia.

METHOD

This study employed an ex-post-facto evaluation method to assess the quality of grade 12 school final tests, a requirement for senior high school graduation in Palembang. Twenty teachers specializing in Islamic education participated, each randomly selecting 40 student response

papers, totaling 800 responses. The evaluation criteria included validity, test reliability, item difficulty, discrimination index, and disruptive impact, drawing from established methodologies (Tuckman, 1985; Secolsky & Denison, 2017).

Teachers in the Islamic education subject were chosen purposefully, ensuring a diverse representation. The selected teachers then randomly picked 40 student response papers from their grade 12 classes, forming a robust dataset of 800 student answers. The focus was on science and social class grade 12 exams, and no alterations were made to the exam conditions.

The quality assessment of the summative exams centered on several critical indicators: validity (correlation between items and total scores), test reliability (internal consistency), item difficulty (proportion of students finding items too easy or complex), discrimination index (performance difference among students), and disruptive impact of distractors in student choices.

The data collection involved document analysis of student response sheets, emphasizing science and social class grade 12 exams. No interventions were made during the evaluation process. Quantitative data were collected after testing study instruments, utilizing ANATES version 4 for a comprehensive investigation of validity, reliability, item difficulty, discrimination index, and distractor efficacy.

FINDING AND DISCUSSION

Finding

Item validity

In this study, item validity was seen from the correlation value between the item and total scores (item total). For the ANATES program, the acceptable range of correlation values for the 50-item test was 0.273 to 1.00 ($\propto = 0.05$). This analysis was performed on all test sets constructed by teachers for Islamic education. Examples of ANATES results for item-total correlation for Teacher 1 (Respondent 1) who taught Islamic education. Based on the correlation of the item score and the total item score, the analysis reveals that only 14 out of the 50 items, or 28% of the total, are valid items for Teacher 1 (Respondent 1). The analysis was repeated for all 20 Islamic education teachers.

The analysis of item validity in this study, focusing on the correlation between individual items and total scores using the ANATES program, aligns with established practices in educational measurement (Van der Linden, 2017). The acceptable correlation range of 0.273 to 1.00, as specified for the 50-item test, is consistent with the importance of evaluating the relationship between individual items and the overall test performance (Gorham & Randall, 2022). Examining the ANATES results for Teacher 1 (Respondent 1) in Islamic education revealed a notable concern. Only 14 out of 50 items, accounting for 28%, demonstrated valid correlations with the total score. This outcome raises questions about the effectiveness of the test items in accurately reflecting students' overall understanding of the subject matter. Similar analyses across all 20 Islamic education teachers can provide a comprehensive understanding of the overall validity of the items and guide potential improvements for future assessments (Livingston & Zieky, 1982).

Reliability test

The reliability of a test refers to the consistency of the test results. In this study, reliability was determined based on internal consistency. The reliability of the test was determined by measuring the depth of its consistency. Internal consistency was analyzed using Cronbach's alpha measurement method, with the category achieving high-reliability levels above 0.71 and a moderate level of .41 to .70, while a poor level was less than 0.40 (Wagemaker, 2020). The instrument used is reliable if the teacher-constructed test has a reliability coefficient of 0.60 (Wyatt-Smith & Adie, 2018). A higher reliability index gives the impression that the measuring instrument is more consistent and can measure a concept accurately (Marzano et al., 2018).

For the subject of Islamic education, eight teachers could construct a test with high reliability. In comparison, ten teachers were able to construct a test with moderate reliability, and

two teachers were able to construct a test with a poor level of reliability. A summary of the teachers' reliability and level of achievement is shown in Table 1.

Table 1. Summary of ANATES results for the reliability of the Islamic education test

Resp.	Reliability Cronbach Alpha	Level of Achievement
R8	0,068056	High
R16	0,065972	High
R10	0,065972	High
R9	0,058333	High
R13	0,057639	High
R5	0,054861	High
R6	0,054861	High
R14	0,049306	High
R18	0,048611	Moderate
R15	0,045139	Moderate
R1	0,044444	Moderate
R3	0,04375	Moderate
R7	0,043056	Moderate
R12	0,042361	Moderate
R19	00.56	Moderate
R11	00.55	Moderate
R2	00.53	Moderate
R4	00.52	Moderate
R17	00.26	Poor
R20	00.21	Poor

Note:
*N = 20
*Number of UAS items = 50 items; tests based on internal consistency
*Level of achievement to measure test reliability: high over 0.71, Moderate 0.41 to 0.70, and Poor less than 0.40.

The assessment of test reliability through internal consistency, as conducted in this study using Cronbach's alpha, is a well-established practice in educational measurement (DeVellis & Thorpe, 2021). The classification of high reliability above 0.71, moderate reliability between 0.41 and 0.70, and poor reliability below 0.40 aligns with commonly accepted standards in psychometrics (Cortina, 1993). According to the ANATES results for Islamic education tests, eight teachers achieved high reliability, ten demonstrated moderate reliability, and two exhibited poor reliability. This distribution highlights variations in the consistency of test results among teachers, emphasizing the importance of addressing factors that may contribute to lower reliability, such as unclear item wording or inadequate coverage of the curriculum (Haladyna & Downing, 2011). The table presents a clear overview of each teacher's reliability level, providing valuable insights for targeted interventions to enhance the overall reliability of teacher-constructed tests.

Difficulty index

According to classical test theory, a poor level of difficulty indicates that the test items are too difficult for the group of students taking the test. In contrast, the high difficulty level indicates that the items are easy for that group of students. According to Wagemaker (2020), the value of the difficulty index for Good is p = 0.40 to 0.69. Although the difficulty index does not determine the quality of the item, there is a relationship between the discrimination index and the difficulty index. A moderate item difficulty index usually has a good discrimination index value. The item analysis showed the percentage of items with a moderate or acceptable difficulty index for each teacher-made test. The resultsare summarized in Table 2.

Table 2. Summary of ANATES results for the Islamic education difficulty index

Resp.	Difficulty Index		Level of Achievements
	No.	%	
R5	24	48	Moderate
R8	24	48	Moderate
R1	22	44	Moderate
R16	21	42	Moderate
R10	21	42	Moderate
R13	21	42	Moderate
R15	21	42	Moderate
R6	18	36	Poor
R18	19	38	Poor
R17	16	32	Poor
R2	16	32	Poor
R19	16	32	Poor
R14	15	30	Poor
R3	15	30	Poor
R11	15	30	Poor
R9	14	28	Poor
R4	14	28	Poor
R12	12	24	Poor
R20	11	22	Poor
R7	9	18	Poor

Note:
*N = 20
*Number of UAS items = 50 items.
*Level of achievement:

- Outstanding = 81%- 100% items with Moderate p
- Good = 61% - 80% items with Moderate p
- Moderate = 41% - 60% items with Moderate p
- Poor = <40% items with Moderate p

This study found that 13 teachers achieved apoor level in constructing items in terms of difficulty level, and seven teachers constructed items with a moderate achievement level (Rahman et al., 2022). The findings also show that no Islamic Education teachers achieved well in constructing items based on the percentage of items with an acceptable difficulty level (Wyatt-Smith & Adie, 2018).

The analysis of the difficulty index in this study aligns with classical test theory, indicating the level of challenge posed by test items for the group of students. As per Gorham and Randall (2022), a difficulty index ranging from 0.40 to 0.69 is considered good, reflecting an appropriate level of challenge for students. The study's findings, as presented in Table 2, reveal that most teachers achieved moderate difficulty in constructing items, with none reaching the good level. This suggests that, on average, the test items were moderately challenging for the students, with room for improvement to achieve an optimal balance between difficulty and discriminative power (Sinharay et al., 2011). The correlation between the discrimination index and the difficulty index, as noted in classical test theory, emphasizes the need to consider both aspects in test construction to ensure the effectiveness of assessments (Crocker & Algina, 1986).

Moreover, the study's identification of 13 teachers with items at a poor difficulty level and seven at a moderate level underscores the variability in item construction quality among teachers. This variation may be attributed to factors such as the clarity of item wording, alignment with instructional objectives, and consideration of students' cognitive levels (Downing, 2006). The absence of teachers achieving a good difficulty level highlights an area for targeted professional development to enhance the quality of item construction in teacher-made tests.

Discrimination index

The discrimination index (D) indicates the ability of the item to differentiate the ability of different students to answer the item correctly. In the classical theory test, the discrimination index was calculated by the percentage difference of students from the high and poor achievement groups who answered the item correctly. Good items have a discrimination index of > .50 to 1.0 (Wagemaker, 2020; Nengsih et al., 2022).

Based on the 50 items constructed, the number and percentage of items with the good level of achievement for the discrimination index received 61 to 80 percent for each teacher. In this study, it was found that the test with a good discrimination index percentage was constructed by one teacher only; a total of 18 teachers were in the poor achievement category, i.e., had a small percentage of items with an acceptable discrimination index, and another teacher was at moderate level. It can be concluded that teachers have not been able to construct items with a good discrimination index. This summary is shown in Table 3.

Table 3. Summary of ANATES results for the Islamic Education Discrimination Index

Resp.	Discrimination Index		Level of Achievements
	No.	%	
R8	34	68	Good
R5	21	42	Moderate
R6	20	40	Poor
R10	19	38	Poor
R13	17	34	Poor
R12	14	28	Poor
R9	13	26	Poor
R16	12	24	Poor
R15	11	22	Poor
R3	11	22	Poor
R7	10	20	Poor
R4	10	20	Poor
R17	8	16	Poor
R19	8	16	Poor
R18	7	14	Poor
R2	6	12	Poor
R1	6	12	Poor
R14	5	10	Poor
R11	4	8	Poor
R20	1	2	Poor

Note:
* N = 20
* Number of UAS items = 50 items
* Level of achievement for the discrimination index:
  - Outstanding = 81%- 100% of items with D are accepted
  - Good = 61% - 80% of items with D are accepted
  - Moderate = 41% - 60% of items with D are accepted
  - Poor = <40% of items with D are accepted

The discrimination index analysis in this study, as presented in Table 3, provides insights into the effectiveness of items in distinguishing between students of varying abilities. According to classical test theory, a good discrimination index falls within the range of > .50 to 1.0 (Gorham & Randall, 2022). The findings reveal a notable challenge, with only one teacher achieving a good discrimination index, while the majority of teachers (18), fall into the poor achievement category. This suggests a widespread difficulty among teachers in constructing items that effectively discriminate between students based on their ability levels.

The limited success in achieving a good discrimination index could be attributed to various factors, including the clarity of item wording, alignment with instructional objectives, and the cognitive demands placed on students (Haladyna & Downing, 2011). It emphasizes the need for teachers to critically evaluate and refine their item construction practices enhancing the discriminatory power of assessments. Additionally, professional development opportunities focused on item analysis and discrimination index improvement may be beneficial for teachers (Linn, 2008). The findings underscore the importance of continuous efforts to improve the quality of teacher-created test items, as a good discrimination index is crucial for generating meaningful insights into students' varying levels of proficiency.

Disruptive effectiveness

The effectiveness of disruptors was seen based on the proportion of students who chose each alternative or option for each item. To test this, Secolsky and Denison (2017) argue that harassers should be removed or reviewed if no candidate chooses them. Disruptors are considered functional if they are selected by the students taking the test. More students should choose good disruptors from the lower group (poor) than the upper group (outstanding) students.

Based on the 50 items constructed, the number and percentage of items with acceptable disruptive effectiveness selected by the students for each teacher were calculated. The summary is shown in Table 4. Regarding Islamic education, only two teachers constructed items with outstanding degrees of disruptive effectiveness, while five were rated as moderate. In addition, this study also showed that 13 teachers were at a poor level of constructing items with disruptive effectiveness. The results of the ANATES for the disruptive index of Islamic education subjects can be shown in Table 4.

Table 4. Summary of ANATES results for the Index of Islamic Education Disruptors

Resp.	Disruptive Effectiveness		Level of Achievements
	No.	%	
R1	42	84	Outstanding
R5	32	64	Good
R6	29	58	Moderate
R2	28	56	Moderate
R8	28	56	Moderate
R7	26	52	Moderate
R10	22	44	Moderate
R15	20	40	Poor
R16	18	36	Poor
R9	17	34	Poor
R14	17	34	Poor
R17	17	34	Poor
R4	17	34	Poor
R3	16	32	Poor
R12	14	28	Poor
R18	14	28	Poor
R13	8	16	Poor
R19	8	16	Poor
R20	4	8	Poor
R11	3	6	Poor

Note:
* N = 20
* Number of UAS items = 50 items;
* Level of achievement to measure the effectiveness of disruptors:
  - Outstanding = 81%- 100% items with intruders accepted

- Good = 61% - 80% items with intruders accepted
- Moderate = 41% - 60% items with intruders accepted
- Poor = <40% items with intruders accepted

Most items are not all distractors work with a good level of disruptive effectiveness. For example, for Respondent 1, an Islamic education teacher, out of 50 UAS test items, only five had a good level of disruptive effectiveness.

The analysis of disruptive effectiveness, as presented in Table 4, sheds light on the performance of distractors in Islamic Education test items. According to Secolsky and Denison (2017), the functionality of distractors is crucial for assessing their effectiveness. An optimal scenario involves all distractors being chosen by students, with a higher selection rate for good distractors by lower-achieving students compared to outstanding ones.

However, the findings indicate a considerable challenge to achieving this ideal scenario. Only two teachers constructed items with outstanding disruptive effectiveness, and five teachers achieved a good level. Conversely, 13 teachers were at a poor level in constructing items with disruptive effectiveness. This suggests a widespread difficulty among teachers in creating distractors that effectively challenge students across different proficiency levels.

To address this challenge, teachers could benefit from professional development opportunities focused on enhancing item construction, particularly in devising effective distractors. Strategies may include refining distractor wording, aligning distractors with common student misconceptions, and ensuring a balanced difficulty level across all distractors (Tarrant et al., 2009).

Additionally, the examples in Table 5 illustrate that, for Respondent 1, only five out of 50 UAS test items had disruptive items that worked well with a good level of effectiveness. This emphasizes the need for teachers to critically evaluate and improve the quality of their distractors, as effective distractors contribute to the overall diagnostic power of assessments (Downing, 2006).

Table 5. Examples of disruptors that work well in Islamic education subjects (R1)

No. Item	Options				
	a	b	c	D	e
7	4+	9+	13**	5+	9+
20	7+	0**	14+	6+	13+
37	16**	4+	8+	4+	8+
43	19**	7+	4++	4++	6++
46	2+	4+	28**	2+	4+

Note:
Number of students = 40 people
Number of Items = 50
**: Answer Key
++: Very Good
+ : Good

Summary of the quality of tests made by teachers of Islamic education subjects

The quality of the teacher-made summative test in this study was assessed based on the percentage of valid items (item-total correlation), the internal consistency indicated by Cronbach's alpha, the rate of items with an acceptable discrimination index (D), as well as the percentage of items with acceptable disruptive effectiveness (Handrianto et al., 2023). The quality of the summative test (UAS) of Islamic education is generally moderate. To find out all 20 teachers' levels of achievement, Islamic education teachers constructed items on the quality of summative tests. This summary is shown in Table 6.

Table 6. Summary of ANATES results for the quality of summative tests of Islamic education

Resp.	Item Validity		Level of Achievements	Cronbach's Alpha (Reliability)	Difficulty Index		Discrimination Index		Disruptive Effectiveness	
	No.	%			No.	%	No.	%	No.	%
R8	46	92	Outstanding	0,0680556	24	48	28	56	34	68
R10	41	82	Outstanding	0,0659722	21	42	22	44	19	38
R16	40	80	Good	0,0659722	21	42	18	36	12	24
R13	30	60	Moderate	0,0576389	21	42	8	16	17	34
R5	27	54	Moderate	0,0548611	24	48	32	64	21	42
R12	27	54	Moderate	0,0423611	12	24	14	28	14	28
R6	26	52	Moderate	0,0548611	18	36	29	58	20	40
R9	24	48	Moderate	0,0583333	14	28	17	34	13	26
R18	24	48	Moderate	0,0486111	19	38	14	28	7	14
R15	24	48	Moderate	0,0451389	21	42	20	40	11	22
R19	23	46	Moderate	00.56	16	32	8	16	8	16
R14	21	42	Moderate	0,0493056	15	30	17	34	5	10
R3	21	42	Moderate	0,04375	15	30	16	32	11	22
R7	21	42	Moderate	0,0430556	9	18	26	52	10	20
R2	18	36	Poor	00.53	16	36	28	56	6	12
R17	17	34	Poor	00.26	16	36	17	34	8	16
R4	17	34	Poor	00.52	14	28	17	34	10	20
R1	14	28	Poor	0,0444444	22	44	42	84	6	12
R11	14	28	Poor	00.55	15	30	3	6	4	8
R20	12	24	Poor	00.21	11	22	4	8	1	2

Note:
* Number of UAS items = 50 items;
* Level of achievement for measuring the quality of Anates results:
  - Outstanding = 81%- 100% accepted items
  - Good = 61% - 80% accepted items
  - Moderate = 41% - 60% accepted items
  - Poor = <40% accepted items

According to the item-total correlation, 46 out of 50 items (92%) had valid items, indicating that Respondent 8 had produced a high-quality test. High (0.98) test reliability was also discovered. However, when comparing the percentage of items with an acceptable discrimination index (68%) to the percentage of items with effective harassment coexistence (56%), the percentage of tests with good quality was lower than 92%.

Specifically, Respondent 16 also showed outstanding test quality, with 40 items (80%) out of 50 test items being legitimate when the number of valid items was evaluated using the correlation between the item score and the total score. Cronbach's alpha value also reached a very high level of 0.95. However, because the discriminating index was still low at just 12 items (24%) and the item for which all harassers chose the students had 18 items (36%), the effectiveness of the harasser was still at a moderate level. This means that teachers cannot construct items with a good discrimination index. This can also be seen from the students 'answer option choices in the distractor analysis (Handrianto et al., 2021).

Another thing about Respondent 11's level of achievement was the poor construction of summative test items. For Respondent 11, valid items were assessed based on the correlation between item scores and total item scores of only 14 items (28%) out of a total of 50 items, with a consistency level of 0.55 on the Cronbach's alpha reliability index. Moreover, in disruptive effectiveness, three items (6%) out of 50 total test items and items with discrimination index are

indicated by calculating the percentage difference between students from high and poor achievement groups who answered the item correctly (poor achievement level was 4 items (8%) only).

For Respondent 20 who constructed poor achievement level items in making summative tests, valid items were assessed based on the correlation between item score and total item score for only 12 items (24%), Cronbach's alpha reliability index 0.21 (poor consistency level, particularly regarding the disruptive effects of 4 items (8%) out of 50 items on the entire test item), and items with a poor 1 item achievement level discrimination index that only hindered students' ability to learn.

CONCLUSION

The study delved into the assessment of summative tests generated by senior high school teachers in Palembang, Indonesia, specifically focusing on the Islamic education subject. The examination encompassed various dimensions, including item validity, reliability, difficulty index, discrimination index, and disruptive effectiveness, revealing a spectrum of test quality among teachers. While some exhibited outstanding test quality, others were categorized as moderate or poor, underscoring the diverse proficiency levels in constructing high-quality summative exams. The study emphasizes the critical importance of addressing this variability in test quality among teachers, as high-quality summative tests play a pivotal role in accurately gauging student achievement and shaping educational decisions. Teachers with exemplary test quality positively contribute to student learning outcomes. In contrast, those with moderate or poor quality may benefit from targeted support and training to refine their test construction skills. The study's insights hold significance for educational institutions and policymakers, guiding the development of interventions to enhance overall test quality and, consequently, improve the precision of student assessments. Despite its contributions, the study acknowledges limitations, such as the subject-specific focus on Islamic education in Palembang, potentially limiting generalizability. Additionally, the study did not explore specific teaching methodologies in test construction, suggesting avenues for future research to delve deeper into these factors influencing test quality. Suggestions for future studies include exploring methodologies, considering subject and regional variations, evaluating the effectiveness of teacher training programs, and conducting longitudinal studies to track the progression of teachers' test construction skills over time. In conclusion, the study underscores the need for targeted interventions to enhance test quality and, by extension, improve the accuracy of student assessments, emphasizing the dynamic nature of teacher-created summative tests.

REFERENCES

Armiati, A., Subhan, M., Nasution, M. L., Al Aziz, S., Rani, M. M., Rifandi, R., & Harisman, Y. (2020). Profesionalisme guru dalam membuat soal higher order thinking skills. *JNPM (Jurnal Nasional Pendidikan Matematika)*, *4*(1), 75-84. http://dx.doi.org/10.33603/jnpm.v4i1.2587

Blakemore, H. (2012). Emergent teacher-researchers: A reflection on the challenges faced when conducting research in the English classroom. *English teaching: Practice and Critique, 11*(2), 59–69.

Chan, K. K. (2018). The effect of teachers' perceptions on the role of technology in assessment: The case of Macau. *International Journal of Learning, Teaching and Educational Research, 17*(2), 127–137.

Cobbold, C., & Wright, L. (2021). Use of Formative feedback to enhance summative performance. *Anatolian Journal of Education*, *6*(1), 109–116.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, *78*(1), 98-104.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications*. Sage publications.

Downing, S. M. (2006). Twelve steps for effective test development. *Handbook of test development*, *3*, 25. Routledge

Gorham, A., & Randall, J. (2022). *Classical Test Theory*. Routledge.

Hagen, T. (2020). Towards a more meaningful evaluation of university lecturers. *New Zealand Journal of Educational Studies*, *55*(2), 379–386.

Haladyna, T. M., & Downing, S. M. (2011). Twelve steps for effective test development. In *Handbook of test development* (pp. 17-40). Routledge.

Handrianto, C., Jusoh, A. J., Goh, P. S. C., Rashid, N. A., & Saputra, E. (2021). Teachers` self-efficacy as a critical determinant of the quality of drug education among malaysian students. *Journal of Drug and Alcohol Research. 10*(3).

Handrianto, C., Jusoh, A. J., Rashid, N. A., Imami, M. K. W., Wahab, S., Rahman, M. A., & Kenedi, A. K. (2023). Validating and testing the teacher self-efficacy (TSE) scale in drug education among secondary school teachers. *International Journal of Learning, Teaching and Educational Research*, *22*(6), 45-58. https://doi.org/10.26803/ijlter.22.6.3

Herawati, N. (2021). Kemampuan guru dalam membuat soal HOTS dalam ujian tengah semester. *Primary: Jurnal Pendidikan Sekolah Dasar*, *10* (6), 1689-1694. http://dx.doi.org/10.33578/jpfkip.v10i6.8638

Herlina, S., Rahman, M. A., Nufus, Z., Handrianto, C., & Masoh, K. (2021). The development of students' learning autonomy using tilawati method at a madrasatul quran in south Kalimantan. *Jurnal Pendidikan Agama Islam*, *18*(2), 431-450. https://doi.org/10.14421/jpai.2021.182-12

Jusuf, R., Sopandi, W., Wulan, A. R., & Sa'ud, U. S. (2019). Strengthening teacher competency through ICARE approach to improve literacy assessment of science creative thinking. *International Journal of Learning, Teaching and Educational Research*, *18*(7), 70-83.

Kartowagiran, B., Wibawa, E. A., Alfarisa, F., & Purnama, D. N. (2019). Can student assessment sheets replace observation sheets? *Jurnal Cakrawala Pendidikan*, *38*(1), 33–44.

Koswara, D., Dallyono, R., Suherman, A., & Hyangsewu, P. (2021). The analytical scoring assessment usage to examine Sundanese students' performance in writing descriptive texts. *Cakrawala Pendidikan*, *40*(3), 573-583.

Laverty, J. T., Bauer, W., Kortemeyer, G., & Westfall, G. (2012). Want to reduce guessing and cheating while making students happier? Give more exams! *The Physics Teacher*, *50*(9), 540-543.

Linn, R. L. (2008). *Measurement and assessment in teaching*. Pearson Education India.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests.

Malone, M. E. (2010). Test review: Canadian academic English language (CAEL) assessment. *Language Testing*, *27*(4), 631–636.

Marzano, R. J., Norford, J. S., & Ruyle, M. (2018). *The new art and science of classroom assessment*. Solution Tree. 555 North Morton Street, Bloomington, IN 47404.

Moeti, B. (2016). Perceptions of teacher counsellors on assessment of guidance and counselling in secondary schools. *International Journal of Learning, Teaching and Educational Research, 15*(6), 145–155.

Mokshein, S. E., Ishak, H., & Ahmad, H. (2019). The use of Rasch measurement model in English testing. *Jurnal Cakrawala Pendidikan, 38*(1), 16-32.

Nengsih, Y. K., Handrianto, C., Nurrizalia, M., Waty, E. R. K., & Shomedran, S. (2022). Media and resources development of android based interactive digital textbook in nonformal education. *Journal of Nonformal Education, 8*(2), 185-191. https://doi.org/10.15294/jne.v8i2.34914

Rahman, M. A., Novitasari, D., Handrianto, C., & Rasool, S. (2022). Assessment challenges in online learning during the covid-19 pandemic. *Kolokium Jurnal Pendidikan Luar Sekolah, 10*(1). https://doi.org/10.24036/kolokium.v10i1.517

Ridho, U. (2018). Evaluasi dalam pembelajaran bahasa Arab. *An Nabighoh, 20*(01), 19-26. https://doi.org/10.32332/an-nabighoh.v20i01.1124

Secolsky, C., & Denison, D. B. (Eds.). (2017). *Handbook on measurement, assessment, and evaluation in higher education*. Routledge.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. Educational Measurement: Issues and Practice, 30(3), 29-40.

Supriyadi, E., Zamtinah, Z., Soenarto, S., & Hatmojo, Y. I. (2019). A character-based assessment model for vocational high schools. *Jurnal Cakrawala Pendidikan, 38*(2), 269-280.

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), 1-8.

Tuckman, B. W. (1985). *Evaluating instructional programs*. Allyn and Bacon, Inc., Rockleigh, NJ 07647.

Van der Linden, W. J. (Ed.). (2017). *Handbook of Item Response Theory: Volume 3: Applications*. CRC press.

Wagemaker, H. (2020). *Reliability and validity of international large-scale assessment: Understanding IEA's comparative studies of student achievement* (p. 277). Springer Nature.

Widiana, I. W., Tegeh, I. M., & Artanayasa, I. W. (2021). The project-based assessment learning model that impacts learning achievement and nationalism attitudes. *Jurnal Cakrawala Pendidikan, 40*(2), 389-401.

Wyatt-Smith, C., & Adie, L. (2018). *Innovation and accountability in teacher education*. Springer Singapore.