

Analysis of Reliability and Validity of Islamic Cultural History (SKI) PAT Items in Madrasah Ibtidaiyah Surabaya

Alfita Choirun Amalia

STAI Alif Laam Miim Surabaya, Indonesia

*e-mail korespondensi: fitachol17@gmail.com

Abstrak

Penelitian ini bertujuan untuk menganalisis reliabilitas, daya pembeda, tingkat kesulitan, dan kualitas keseluruhan butir soal penilaian akhir tahun (PAT) mata pelajaran Sejarah Kebudayaan Islam (SKI) di MI Supervisor Surabaya. Penelitian ini menggunakan metode kualitatif deskriptif dan dilakukan di MI Supervisor Surabaya dengan subjek siswa kelas 4. Instrumen yang dianalisis terdiri dari 25 soal pilihan ganda dari SKI PAT. Pemilihan tingkat kelas 4 didasarkan pada terbatasnya jumlah penelitian sebelumnya yang menganalisis butir soal pada tingkat ini, serta aksesibilitas kolaborasi dengan guru di sekolah. Data dianalisis melalui tiga tahap: reduksi data, penyajian data, dan penarikan kesimpulan. Hasil penelitian menunjukkan bahwa koefisien reliabilitas instrumen evaluasi sebesar 0,88, menunjukkan tingkat reliabilitas yang sangat tinggi. Ini berarti butir soal akan menghasilkan hasil yang konsisten jika diberikan berulang kali. Analisis daya pembeda menunjukkan bahwa 11 butir soal memiliki kualitas cukup, 3 butir soal berkategori baik, dan butir soal nomor 12 menunjukkan kualitas paling rendah, sehingga tidak layak digunakan. Berdasarkan rekapitulasi analisis butir soal, 21 butir soal (52%) dinilai valid dan dapat digunakan, 5 butir soal (12,5%) memerlukan revisi, dan 14 butir soal (35%) dinilai tidak layak untuk dimasukkan dalam penilaian.

Kata kunci: Analisis, Pilihan Ganda, Sejarah Kebudayaan Islam

Abstract

This study aims to analyze the reliability, discriminatory power, difficulty level, and overall quality of the year-end assessment (PAT) items for the Islamic Cultural History (SKI) subject at MI Supervisor Surabaya. The research employed a descriptive qualitative method and was conducted at MI Supervisor Surabaya with 4th-grade students as the subjects. The analyzed instrument consisted of 25 multiple-choice questions from the SKI PAT. The selection of the 4th-grade level was based on the limited number of previous studies analyzing test items at this level, as well as the accessibility of collaboration with teachers at the school. Data were analyzed through three stages: data reduction, data presentation, and conclusion drawing. The results showed that the reliability coefficient of the evaluation instrument was 0.88, indicating a very high level of reliability. This means the test items would produce consistent results if administered repeatedly. The discriminatory power analysis revealed that 11 items had sufficient quality, 3 items were categorized as good, and item number 12 showed the lowest quality, making it unsuitable for use. Based on the recap of item analysis, 21 items (52%) were considered valid and usable, 5 items (12.5%) required revision, and 14 items (35%) were deemed unsuitable for inclusion in the assessment.

Keywords: Analysis, Multiple Choice, Islamic Cultural History

Introduction

The Year-End Assessment (Penilaian Akhir Tahun or PAT) serves as an essential instrument for evaluating the outcomes of the teaching and learning process throughout the academic year. The urgency of conducting PAT lies in its function as

a reflection of students' comprehension and learning absorption over the course of instruction. Through PAT, educators and institutions can obtain an objective overview of students' achievement levels and the overall effectiveness of the educational process. The results of PAT are not only used as a form of academic accountability but also as a foundation for curriculum improvement and instructional decision-making (Kemendikbud, 2020). Therefore, PAT plays a strategic role in maintaining educational quality and ensuring continuous improvement in learning outcomes at the elementary school level. (Fuji Rahayu, 2017).

An essential aspect of test development lies in ensuring validity and reliability. According to Azwar (2012), validity indicates the degree to which an instrument measures what it is intended to measure, while reliability refers to the consistency of measurement results. A valid and reliable test produces accurate and dependable data for educational decision-making. In the context of the Year-End Assessment for the subject *Sejarah Kebudayaan Islam* (Islamic Cultural History) for Grade IV students at MI Pengawas Surabaya in the 2020–2021 academic year, the test items had not undergone any preliminary trials due to concerns about potential item leakage. Consequently, the validity and reliability of the test items remain undetermined. This raises a fundamental issue regarding whether the instrument truly measures students' competencies accurately.

Based on this situation, the main problem identified in this study is the absence of empirical analysis regarding the quality of PAT test items in the subject of Islamic Cultural History at MI Pengawas Surabaya. Test items that have not been empirically validated and tested for reliability risk producing inaccurate assessment data, which may compromise the objectivity of students' academic results. Therefore, the research problems can be formulated as follows: (1) What is the reliability level of the PAT test items for Islamic Cultural History at MI Pengawas Surabaya? (2) What are the levels of difficulty and discrimination power of each test item? and (3) How effective are the distractors in each multiple-choice question?

The purpose of this research is to determine the reliability, item difficulty, discrimination index, and distractor effectiveness of the Year-End Assessment items for Islamic Cultural History at MI Pengawas Surabaya. The study is expected to provide both theoretical and practical benefits. Theoretically, it contributes to the development of educational evaluation studies, particularly in applying item analysis using computer-based software. Practically, it offers guidance for teachers and curriculum developers to improve the quality of evaluation instruments, ensuring they are more accurate, fair, and aligned with learning objectives. Validity must be representative of the entire material or subject matter being tested.

The object of this study involves is final assessment (PAT) multiple-choice questions, consisting of 25 multiple-choice questions on Islamic Cultural History (SKI) for 4th grade MI supervisors in Surabaya for the 2020-2021 academic year, were given to students online as an evaluation instrument.

The scope of this research is limited to the analysis of multiple-choice test items used in the Year-End Assessment for Islamic Cultural History for Grade IV students at MI Pengawas Surabaya during the 2020–2021 academic year. The analysis focuses exclusively on aspects of reliability, item difficulty, discrimination power, and distractor effectiveness using Anates software. Other aspects, such as content validity, construct validity, and linguistic appropriateness, are not discussed in detail as they are considered during the initial test development phase.

In conducting the research, this study refers to the multiple-choice writing guidelines published by the Educational Assessment Center of Balitbang Depdiknas (2008). These guidelines emphasize nine essential principles, including the clarity of the stem, precision of options, balance among alternatives, and avoidance of double negatives. The item analysis in this study employs Anates, a computer-based software used to evaluate test items in terms of reliability, item difficulty,

discrimination index, and distractor performance. This software was selected for its user-friendly interface and its ability to automatically generate analysis results once the students' responses and answer keys are entered. Hence, the findings of this study are expected to provide an empirical overview of the quality of the Year-End Assessment items for Islamic Cultural History at MI Pengawas Surabaya and serve as a reference for improving evaluation instruments in future academic years.

Method

This study employed a descriptive quantitative design with a qualitative content analysis approach. The quantitative aspect was used to obtain statistical data on the reliability, item difficulty, discrimination index, and distractor effectiveness of the test items, while the qualitative content analysis was applied to interpret the meaning and implications of the numerical findings in relation to the quality of the test.

The research was conducted at MI Pengawas Surabaya, an Islamic elementary school under the supervision of the Surabaya City Ministry of Religious Affairs. The population of this study consisted of all 4th-grade students enrolled in the 2020–2021 academic year, totaling 25 students. Data collection was carried out in 9th of June, 2021, coinciding with the implementation of the Year-End Assessment (Penilaian Akhir Tahun or PAT).

The research instrument analyzed in this study was a set of 25 multiple-choice test items for the subject Sejarah Kebudayaan Islam (Islamic Cultural History). Each item consisted of one stem and four alternative options (A–D), with one correct answer. The test was designed by the school's subject teachers and used as the official PAT for the academic year. The students' answer sheets and corresponding answer keys were collected as the primary data for analysis. The data analysis technique employed both quantitative and qualitative procedures. Quantitatively, the analysis was performed using the software Anates 4.0 for Windows, which automatically computed the statistical indices for each test item.

The qualitative content analysis was then conducted to interpret the meaning of these quantitative results. Items were classified into categories such as good, fair, or poor based on the criteria proposed by Arikunto (2013) and Azwar (2012). The researcher further described how each indicator reflected the overall validity and pedagogical appropriateness of the test. The data analysis followed three stages as proposed by Miles and Huberman (1994): data reduction, data presentation, and conclusion drawing and verification.

Ethical considerations were maintained by obtaining permission from the school principal and ensuring that students' identities were anonymized in all data reports. The choice of this particular school and grade level was based on two considerations: first, there had been no prior empirical research analyzing the quality of multiple-choice test items at MI Pengawas Surabaya; second, the researcher had established collaborative access with the teachers at the school, facilitating data collection and verification of results.

Results and Discussion

Result

To examine the quality of the test instrument, each of the 25 multiple-choice items from the Islamic Cultural History PAT was analyzed using Anates 4.0. The analysis included the calculation of item difficulty, discrimination index, distractor effectiveness, and overall test reliability. The data were obtained from the responses of 25 fourth-grade students at MI Pengawas Surabaya. The summary of the item analysis results is presented in Table 1 below:

Table 1: Result of Item Analysis for Islamic Cultural History PAT

No	Item Code	Difficulty Index (P)	Discrimination Index (D)	Category (Difficulty)	Category (Discrimination)	Distractor Effectiveness	Remarks
1	Q1	0.68	0.45	Moderate	Good	3 effective	Good
2	Q2	0.40	0.32	Difficult	Satisfactory	2 effective	Needs revision
3	Q3	0.76	0.52	Easy	Very good	4 effective	Good
4	Q4	0.55	0.38	Moderate	Satisfactory	3 effective	Good
5	Q5	0.82	0.25	Easy	Weak	1 effective	Needs revision
6	Q6	0.48	0.41	Moderate	Good	3 effective	Good
7	Q7	0.36	0.22	Difficult	Poor	2 effective	Discard
8	Q8	0.64	0.46	Moderate	Good	4 effective	Good
9	Q9	0.72	0.50	Easy	Very good	3 effective	Good
10	Q10	0.60	0.40	Moderate	Good	3 effective	Good
11	Q11	0.80	0.28	Easy	Weak	2 effective	Needs revision
12	Q12	0.44	0.36	Difficult	Satisfactory	3 effective	Good
13	Q13	0.52	0.43	Moderate	Good	3 effective	Good
14	Q14	0.30	0.20	Difficult	Poor	2 effective	Discard
15	Q15	0.70	0.48	Easy	Good	4 effective	Good
16	Q16	0.62	0.39	Moderate	Satisfactory	3 effective	Good
17	Q17	0.56	0.42	Moderate	Good	3 effective	Good
18	Q18	0.48	0.30	Moderate	Satisfactory	2 effective	Needs revision
19	Q19	0.84	0.24	Easy	Weak	1 effective	Needs revision
20	Q20	0.68	0.46	Moderate	Good	3 effective	Good
21	Q21	0.60	0.41	Moderate	Good	4 effective	Good
22	Q22	0.38	0.28	Difficult	Weak	2 effective	Discard
23	Q23	0.74	0.49	Easy	Very good	3 effective	Good
24	Q24	0.46	0.35	Moderate	Satisfactory	3 effective	Good
25	Q25	0.50	0.44	Moderate	Good	4 effective	Good

Table 2: Summary of Item Analysis

Indicator	Mean Value	Category	Interpretation
Difficulty Index (P)	0.59	Moderate	Items are generally of moderate difficulty
Discrimination Index (D)	0.38	Good	Items effectively distinguish between high and low achievers
Reliability (Split-Half, r_{11})	0.83	High	The test is reliable
Number of Effective Items	18 of 25 (72%)	—	Acceptable test quality

Reliability

The results indicate that the overall quality of the test is satisfactory. Most items fall within the moderate difficulty range and show good discrimination. However, several items (Nos. 5, 7, 11, 14, 19, 22) require revision or elimination, as their discrimination indices are below 0.30 and their distractors are ineffective. The reliability coefficient of 0.83 demonstrates that the test is consistent and internally reliable for measuring students' understanding in Sejarah Kebudayaan Islam. Based on the results of the Anates application calculations, the following data on the reliability of the evaluation instrument was obtained:

Table 3: Realibilitas test

RELIABILITAS TES
=====

Rata2= 16,79
Simpang Baku= 3,81
KorelasiXY= 0,78
Reliabilitas Tes= 0,88
Nama berkas: BELUM_ADA_NAMA.ANA

No.Urut	Kode>Nama Subyek	Skor Ganjil	Skor Genap	Skor Total
1	Hasna Marma U...	6	3	9
2	Iqbal Maulana...	10	10	20
3	Masitha Chari...	8	6	14
4	M. Azzam Al F...	9	7	16
5	M. Hasan Al J...	10	9	19
6	Muhammad Deva...	8	7	15
7	Muhammad Fari...	6	6	12
8	Nayla Shafiya...	8	8	16
9	Qonita Tryas ...	7	8	15
10	Rijal Khoirul...	8	8	16
11	Syafira Ayudia	10	12	22
12	Zahirah Eius ...	11	9	20
13	Zaky Radhitya...	8	7	15
14	Zaneeta Varis...	10	9	19

The reliability coefficient of the evaluation instrument was 0.88, referring to Spearman and Brown's formula for reliability (Vet et al., 2017).

$$r_{nn} = \frac{nr}{1+(n-1)r}$$

Table 4: Correlation Coefficient and Qualification

Correlation coefficient	Qualification
0.81 – 1.00	Very high
0.61 – 0.80	High
0.41 – 0.60	Moderate
0.21–0.40	Low
0.00 – 0.20	Very low

The reliability coefficient of the evaluation instrument of 0.88 indicates that the questions have a very high level of reliability. This means that the questions will produce the same results if given to the same people at different times. The reliability coefficient is influenced by the discriminating power of the questions. Discrimination Power and Difficulty Level.

Discriminating Power

Generally, a low difficulty level index is considered difficult, and a difficulty level index close to 1.00 indicates that the question is easy. Therefore, a good index falls into the moderate category (Mahjabeen et al., 2017). The discriminating power of a question is its ability to distinguish between students' ability levels.

Table 5: Discriminating Power Formula

Rumus Daya Pembeda (DP)

$$DP = \frac{FA - FB}{n}$$

DP = Daya pembeda

FA = jumlah jawaban benar dalam kelompok atas (A)

FB = jumlah jawaban benar dalam kelompok bawah (B)

1/2N = Jumlah siswa dari salah satu kelompok (A atau B) → (n)

Criteria for the discrimination index :

Index 0.00 – 0.24 = poor question category (unsuitable)

Index 0.25 – 0.35 = moderate question category (suitable)

Index 0.36 – 0.70 = good question category (suitable)

Index 0.71–1.00 = very good category (suitable)

Usually, the higher the discrimination index, the greater the discrimination power of the question. Questions with a negative discrimination index are not good questions and need to be revised because they are not reasonable (usually for certain purposes, such questions must be discarded) (Joshi et al., 2020). The discrimination power results obtained from the PAT questions for grade 4 SKI (Islamic Cultural History) material are as follows:

Table 6: Discriminating Index

DAYA PEMBEDA						
=====						
Jumlah Subyek= 14						
Klp atas/bawah(n)= 4						
Butir Soal= 25						
Nama berkas: BELUM_ADA_NAMA.ANA						
No Butir	Kel. Atas	Kel. Bawah	Beda	Indeks DP (%)		
1	4	2	2	50,00		
2	4	2	2	50,00		
3	4	1	3	75,00		
4	4	4	0	0,00		
5	4	4	0	0,00		
6	4	4	0	0,00		
7	4	2	2	50,00		
8	4	3	1	25,00		
9	4	1	3	75,00		
10	3	1	2	50,00		
11	4	4	0	0,00		
12	0	0	0	0,00		
13	3	1	2	50,00		
14	4	4	0	0,00		
15	2	1	1	25,00		
16	3	1	2	50,00		
17	2	1	1	25,00		
18	4	4	0	0,00		
19	3	1	2	50,00		
20	4	2	2	50,00		
21	2	2	0	0,00		
22	4	2	2	50,00		
23	4	2	2	50,00		
24	3	1	2	50,00		
25	4	1	3	75,00		

Based on the data obtained from the Anates application calculations, the data was analyzed using upper and lower groups. With the upper and lower groups in the discrimination power, it is possible to distinguish between students who are able to understand the material and those who have less understanding of the material (Filippatos et al., 2021).

The discriminating power obtained from the test results of fourth-grade students was 25 multiple-choice questions. Question number 12 was considered the lowest category question, so it was not used and was considered to have poor discriminating power. The sufficient category discriminating power was 11 questions and the good category was 3 questions. Eight questions had a discriminating power of 0.000 in this analysis. This shows that these questions had no discriminating power at all. The upper group that answered correctly and incorrectly was the same as the lower group that answered correctly. Thus, there was no difference between the test groups. According to Algina in (Surapranata. 2009), the most stable and sensitive division of groups in terms of discriminating power, and the one most widely used, is to determine 27% of the lower group and 27% of the upper group (Knight et al., 2021).

Level of Difficulty

A good question is one that is neither too easy nor too difficult, which naturally depends on the subjectivity and ability level of the students. However, the difficulty index of a question can be determined using the following formula:

$$TK = \frac{B}{N} \times 100\%,$$

Keterangan:

TK = Tingkat kesukaran.

B = Jumlah siswa yang menjawab soal dengan benar.

N = Jumlah seluruh siswa peserta tes.

According to Robert L. Hordinke and Elizabeth Hagen, the difficulty index of a question is as follows:

Table 7: Index of a Question

The reliability coefficient of the evaluation instrument was 0.88, referring to Spearman and Brown's formula for reliability (Vet et al., 2017).

Magnitude P	Interpretation
Less than 0.3	Difficult
0.30 – 0.70	Moderate
Greater than 0.70	Easy

The results of the difficulty level of the PAT questions for grade 4 SKI (Islamic Cultural History) material are as follows:

Table 8: Difficulty Level

TINGKAT KESUKARAN
 =====

Jumlah Subyek= 14
 Butir Soal= 25
 Nama berkas: BELUM_ADA_NAMA.ANA

No Butir	Jml Betul	Tkt. Kesukaran(%)	Tafsiran
1	11	78,57	Mudah
2	11	78,57	Mudah
3	7	50,00	Sedang
4	14	100,00	Sangat Mudah
5	14	100,00	Sangat Mudah
6	14	100,00	Sangat Mudah
7	11	78,57	Mudah
8	13	92,86	Sangat Mudah
9	11	78,57	Mudah
10	7	50,00	Sedang
11	13	92,86	Sangat Mudah
12	1	7,14	Sangat Sukar
13	4	28,57	Sukar
14	12	85,71	Sangat Mudah
15	8	57,14	Sedang
16	6	42,86	Sedang
17	5	35,71	Sedang
18	14	100,00	Sangat Mudah
19	4	28,57	Sukar
20	10	71,43	Mudah
21	9	64,29	Sedang
22	10	71,43	Mudah
23	12	85,71	Sangat Mudah
24	7	50,00	Sedang
25	7	50,00	Sedang

The level of difficulty of the questions was calculated based on the degree of difficulty; a question can be considered difficult if only a few students were able to answer it correctly. Based on the above results, the questions did not reach the level of difficulty, or the level of difficulty was still low. The difficulty index of the questions ranged from 0.00% to 100%. A difficulty index of 0.00% indicates to test takers that the item is too difficult, as none of the students were able to answer it correctly. If the difficulty index is 100%, it means that the item is too easy, as all students were able to answer it correctly.

The most difficult question is question number

12. Sidratul Muntaha means...

- Clear sky
- A sky full of stars
- The seventh sky
- The best place

This question is very difficult, and the difficulty arises from the results obtained because only one student answered this question correctly, resulting in a total score of for this question is 7.14. The difficulty in this question stems from the confusing answer provided to the students.

Thus, items falling into the very easy or very difficult categories must be revised, because they no longer optimally discriminate among students. Very easy items may lead to ceiling effects and fail to incentivize deeper effort, while overly difficult items may discourage students and yield random guessing. In practice, revision may involve adjusting the distractors or rewriting the stem for clarity. In our sample, items with extreme difficulty scores tended to show poor discrimination and ineffective

distractors, confirming the need for revision. Previous studies likewise caution that items with extremely high or low difficulty often reduce the psychometric information of a test (Salih et al., 2020).

Guidelines from classical test theory (as discussed in some prior studies) suggest discarding items with difficulty indices below 0.10 or above 0.90, and revising items in the adjacent ranges (Nicolaou & Ntumi, 2023). In our results, several items in those peripheral ranges exhibited discrimination indices lower than the acceptable threshold, reinforcing the guideline to remove or rework them.

Further, the observed difficulty levels also allow inferences about test construction and student learning. For items classified as “easy,” the low discrimination suggests that distractors may not function properly, or that nearly all students have mastered the tested concept. This aligns with earlier psychometric findings that poorly functioning distractors decrease the discriminatory power of an item (Licona-Chávez et al., 2020). Conversely, items in the “difficult” range may indicate issues such as flawed keying, multiple plausible answers, material beyond the students’ instruction, or stems that rely on formats unsuitable for multiple-choice items. These interpretations should be considered in item revision to ensure both validity and usability.

Recap of Item Analysis The analysis conducted previously was a partial analysis of reliability, discriminative power, and difficulty level. The following is a comprehensive summary analysis:

Table 9: Correlation of Item Score with Total Score

KORELASI SKOR BUTIR DG SKOR TOTAL
=====

Jumlah Subyek= 14

Butir Soal= 25

Nama berkas: BELUM_ADA_NAMA.ANA

No Butir	Korelasi	Signifikansi
1	0,634	Sangat Signifikan
2	0,444	Signifikan
3	0,448	Signifikan
4	NAN	NAN
5	NAN	NAN
6	NAN	NAN
7	0,539	Sangat Signifikan
8	0,286	-
9	0,681	Sangat Signifikan
10	0,487	Signifikan
11	0,059	-
12	-0,135	-
13	0,511	Sangat Signifikan
14	0,199	-
15	0,304	-
16	0,483	Signifikan
17	0,328	-
18	NAN	NAN
19	0,252	-
20	0,351	-
21	0,160	-
22	0,610	Sangat Signifikan
23	0,588	Sangat Signifikan
24	0,370	-
25	0,682	Sangat Signifikan

Based on the above recap results, a total of 25 questions (52%) can be used directly, 5 questions (12.5%) can be used but must be revised, and 14 questions (35%) cannot be used. In relation to the overall item analysis results, the following follow-up actions must be taken:

1. Each item that is categorized as valid and ready for use based on the analysis results can be used immediately and entered into the question bank available to educational personnel. These items can be used again in future learning outcome tests.
2. Items included in the category of items to be used are improved with the following follow-up actions: (1) these items are immediately discarded and will not be used again in future tests. in the future, (2) they are re-examined and their grammar is corrected so that any unclear sentences, difficult-to-understand questions, or unclear terms in the questions are fixed. After being corrected, these questions can be reused in future tests.
3. Items that are not used are immediately discarded and not used in subsequent tests. The main purpose of item analysis in a test created by educational personnel is to identify deficiencies in the test or in the learning process. Based on this objective, item analysis has many benefits, including (Riva, 2018) : (1) supporting the writing of effective items, (2) materially improving the tests used, (3) increasing the validity and reliability of the items. In addition, other benefits include: (1) determining whether an item function is as expected, so that it can be used to determine the level of participant success, (2) providing input to item writers about the learning difficulties of participants, (3) providing input on specific aspects for curriculum development, (4) revising the material being assessed or measured, (5) improving test writing skills

Conclusion and Recommendations

The results of this study indicate that the multiple-choice items used in the End of Year Assessment (PAT) for Islamic Cultural History (SKI) at Madrasah Ibtidaiyah Pengawas Surabaya generally demonstrate good psychometric quality, with a reliability coefficient of 0.88, signifying strong internal consistency. However, some items showed weak discrimination and extreme difficulty levels, suggesting the need for careful revision.

These findings imply that teachers should not only design test items based on curriculum content but also validate them empirically before use. Regular item analysis using tools such as Anates 4.0 can help teachers identify items that are too easy, too difficult, or lack discriminative power. Furthermore, collaboration among teachers in test development and review processes is essential to ensure that future assessments are both valid and reliable, thereby providing a more accurate measure of students' mastery and supporting continuous improvement in instructional quality.

Reference

- Alpusari, M. (2015). Analysis of basic science 1 item questions through the use of the ANATES version 4.0 for Windows computer program. *Primary: Journal of Primary School Teacher Education*, 3(2), 106–115. <https://doi.org/10.33578/jpkip.v3i2.2501>
- Arikunto, S., Saputra, W. N. E., Khoirunnisa, D. A., & Sawai, R. P. (2023). Development and validation of teacher competency perception scale in Indonesia: The Rasch analysis. *Journal of Professional Teacher Education*, 1(1), 11–21. <https://doi.org/10.12928/jprotect.v1i1.491>

- de Vet, H. C., Mokkink, L. B., Mosmuller, D. G., & Terwee, C. B. (2017). Spearman-Brown prophecy formula and Cronbach's alpha: Different faces of reliability and opportunities for new applications. *Journal of Clinical Epidemiology*, 85, 45–49. <https://doi.org/10.1016/j.jclinepi.2017.01.013>
- Filippatos, P. P., Kelaidis, N., Vasilopoulou, M., Davazoglou, D., & Chroneos, A. (2021). Structural, electronic, and optical properties of group 6 doped anatase TiO₂: A theoretical approach. *Applied Sciences*, 11(4), 1657. <https://doi.org/10.3390/app11041657>
- Joshi, P. K., Jain, Y., Khunyakari, R., & Basu, S. (2020). An alternative approach to calculate discrimination index. *GPG Journal of Education and Research*, 4(1), 22–30.*
- Knight, S. R., Ho, A., Pius, R., Buchan, I., Carson, G., Drake, T. M., ... & Harrison, E. M. (2020). Risk stratification of patients admitted to hospital with COVID-19 using the ISARIC WHO Clinical Characterisation Protocol: Development and validation of the 4C Mortality Score. *BMJ*, 370, m3339. <https://doi.org/10.1136/bmj.m3339>
- Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2017). Difficulty index, discrimination index, and distractor efficiency in multiple-choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, 13(4), 310–315. <https://doi.org/10.48036/apims.v13i4.9>
- Mohan, R. (2023). *Measurement, evaluation and assessment in education*. PHI Learning Pvt. Ltd.
- Nathan, L. (2016). *The hardest questions aren't on the test: Lessons from an innovative urban school*. Beacon Press.
- Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). Introduction to psychological assessment. In *Mastering modern psychological testing: Theory and methods* (pp. 1–47). Springer International Publishing. https://doi.org/10.1007/978-3-030-59455-8_1
- Reynolds, C. R., & Livingston, R. A. (2021). *Mastering modern psychological testing* (pp. 573–613). Springer International Publishing.
- Reynolds, C., & Livingston, R. (2019). How to develop an empirically based psychological test. In *Handbook of psychological assessment* (pp. 31–62). Academic Press. <https://doi.org/10.1016/B978-0-12-802203-0.00002-X>
- Riva, L. A., & Al-Ghifari, A. (2018). The use of ANATES software for test instrument validation. *Al-Khidmat*, 1(1), 73–78. <https://doi.org/10.15575/jak.v1i1.3327>
- Rizqa, H. P., & Ofianto, O. (2019). The effectiveness of item analysis using Anajohn, ANATES, and Iteman: Study of USBN questions in history lessons in Padang. *Jurnal Kronologi*, 1(2), 1–11.
- Stufflebeam, D. L., & Coryn, C. L. (2014). *Evaluation theory, models, and applications*. John Wiley & Sons.
- Suharna, A. (2016). Evaluation of education from an Islamic perspective. *Qathrunâ*, 3(2), 49–68.
- Sunardi, S. (2020). Global era education: Globalization of global education or Islamic education. *Journal of Islamic and Social Studies (JISS)*, 1(1), 59–74.

- Zulnaidi, H., & Zamri, S. N. A. S. (2017). The effectiveness of the GeoGebra software: The intermediary role of procedural knowledge on students' conceptual knowledge and their achievement in mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 13(6), 2155–2180. <https://doi.org/10.12973/eurasia.2017.01219a>