Journal of Proceedings Series No. 3 (2025)
The 3rd International Conference of Islamic Education
July 8th 2025, Institut Islam Mamba'ul 'Ulum Surakarta, Kota Surakarta, Indonesia

Analysis of Tarikh Test Items in Madrasah Aliyah Using the Rasch Model Theory

Joko Subando[1*], Meti Fatimah[2], Sukari[3]

[1*]Institut Islam Mambaul Ulum, Surakarta, Indonesia
jokosubando@yahoo.co.id
[2] Institut Islam Mambaul Ulum, Surakarta, Indonesia
fatimahcan@gmail.com
[3]Institut Islam Mambaul Ulum, Surakarta, Indonesia
sukari@iimsurakarta.ac.id

ARTICLE INFO

Article History:
Recieved : 08-June-2025
Revised : 08-June-2025
Accepted : 21-May-2026
Available online : 21-May-2026

Keyword:
Analysis;
Item;
Rasch Model;

ABSTRACT

This study is an evaluative research aimed at analyzing the quality of Fiqh exam items for 12th-grade students at Madrasah Aliyah Negeri 1 Surakarta, using the Rasch Model evaluation theory approach. The reason for choosing this model is its ability to simultaneously evaluate both the test items and the ability to address the limitations of classical test theory. The research subjects consisted of 332 students who met the minimum criteria for Rasch analysis according to Linacre (1994). Data was collected through documentation of final exam results and analyzed using the Winstep software. The main component evaluated in this study is the difficulty level of the items. The analysis was conducted by considering the validity and reliability of the instrument. Validity was assessed through item fit (Outfit MNSQ, ZSTD, and Pt Mean Corr), unidimensionality, and bias detection using DIF and DPF. The results of the study showed that two items (S17 and S27) were outliers in terms of difficulty, two other items (S17 and S39) were classified as misfits, and 19 items were identified as biased. This indicates that some items were unfair and inaccurate in measuring students' abilities. In terms of reliability, the person reliability value of 0.79 indicated a fair consistency in student responses, while the item reliability of 0.99 suggested that the items were highly reliable in distinguishing levels of ability. The Cronbach's alpha of 0.89 indicated very good internal consistency.

This is an open access article under the CC-BY-SA license

How to Cite:

Subando, J., Fatimah, M., & Sukari. (2025). Analysis of Tarikh Test Items in Madrasah Aliyah Using the Rasch Model Theory. *Internasional Conference on Islamic Education, 2*(1), pp-pp.

https://doi.org/10.

INTRODUCTION

In the context of formal education, learning outcome evaluation serves as a crucial instrument for measuring students' competency achievements (Wisman, Effrata, & Tutesa, 2021). One of the most commonly used evaluation tools is testing in the form of itemized questions. A well-constructed test item should serve as an accurate, objective, and representative measure of student ability (Yusuf, 2023). With the government's plan to reinstate the National Examination (UN) model in 2026 (Rezky, 2025), a major challenge that arises is ensuring the quality of the test items used. Item quality must meet various criteria such as varied difficulty levels, absence of bias, resistance to guesswork, and the ability to assess competencies of students from diverse backgrounds and regions with differing educational standards.

In this context, the need for a high-quality item bank becomes increasingly urgent. An item bank should not merely serve as a question repository, but also as a source of rigorously validated and analyzed items, thus aiding teachers in assembling tests that meet established measurement standards (Widana, 2014). The government, in this case the Ministry of Primary and Secondary Education, needs to collaborate with schools to build a representative item bank—particularly for subjects with specific content requirements, such as Tarikh (Islamic cultural history).

Tarikh is a religious subject with unique characteristics. It not only requires cognitive abilities to recall and understand historical events but also involves affective dimensions and values related to faith and Islamic character. This makes item analysis in Tarikh more complex than in general subjects. Factors such as ideological background, religious understanding, and sensitivity to historical issues in Islam can affect the quality of the items, both in terms of content and how the questions are perceived by students.

Nevertheless, research on item analysis in religious subjects remains limited, especially in Tarikh. Most previous studies have focused on general subjects such as Mathematics, Indonesian Language, or Natural Sciences. A notable study by Saepuzaman, Haryanto, Istiyono, Retnawati, & Yustiandi (2021) analyzed Physics test items on the topic of Work and Energy using both dichotomous and polytomous scoring, as well as BILOG-MG and R software for analysis. Their results showed that most items had good characteristics based on the two-parameter logistic (2PL) model, demonstrating that IRT can provide more detailed insights into item quality than classical approaches.

A similar study by Alfarisa & Purnama (2019) analyzed the characteristics of end-of-semester (UAS) economics test items for 11th-grade students in Yogyakarta using the Rasch model. The study involved 306 students, with data collected via stratified proportional random sampling. Forty multiple-choice questions were validated through expert judgment and Aiken's V index, and had a KR-20 reliability score of 0.77. The results showed that 39 items fit the Rasch model, with difficulty levels ranging from -2.51 to 1.77, and a maximum information function value of 18.3379.

In the field of Natural Sciences, Novriyanti & Arthur (2024) evaluated the quality of midterm exam items for General Biology (Ibn Utsman Al-Faris) using Rasch analysis.

Out of 75 multiple-choice questions, item reliability was very high (0.97). Most items were within the acceptable Infit and Outfit range (0.5–1.5), although some items—such as Item 13 and 54—were identified as misfits. Wright Map analysis indicated good coverage for medium-level abilities, but revealed gaps at the extreme levels.

In Mathematics, Parisu, Ekadayanti, Sisi, & Juwairiyah (2024) analyzed basic mathematics test items for elementary students using a Rasch approach to produce a high-quality test instrument. The study used a sample of 75 students and an instrument consisting of 25 multiple-choice questions, analyzed with Winstep software. The results showed 24 valid items and one invalid item. The difficulty classification included six very difficult, seven difficult, nine easy, and three very easy items, with good reliability assurance for item differentiation and test packaging (0.94).

Subando, Saleh, & Ngatimin (2022) previously conducted an item analysis for Tarikh at Madrasah Quraniyah Al Husnayain Surakarta. However, based on literature review, no study has yet analyzed Tarikh test items at MAN 1 Surakarta. Given that Tarikh is a compulsory subject for madrasah aliyah students, item quality in this subject deserves the same serious attention as that given to other subjects. Good Tarikh items should not only be historically accurate, but also free from ideological bias and capable of fostering critical thinking and deep understanding of Islamic history.

This situation highlights the urgency of research focusing on the quality analysis of Tarikh test items. Such research will not only contribute to improving assessment quality in madrasahs but can also support national efforts to develop a standardized and high-quality item bank. Through systematic analysis based on established measurement theories, teachers and item developers can identify which test items meet the necessary validity and reliability criteria for learning evaluation.

This study focuses on analyzing the quality of Tarikh test items at Madrasah Aliyah Negeri 1 Surakarta. This location was chosen because the school serves as a model for both public and private madrasah aliyah institutions in the Surakarta area. Therefore, analyzing Tarikh items at this school may represent the real challenges in developing high-quality religious test items. Furthermore, the results of this study are expected to offer concrete recommendations to madrasah teachers, particularly in constructing items that are not only technically sound but also reflective of pedagogical and Islamic values. Ultimately, this research aims to contribute to the development of a valid and reliable Tarikh item bank and assist the government in realizing a better national education evaluation system.

LITERATURE REVIEW

Item Response Theory (Rasch Model)

Item Response Theory (IRT) was developed to overcome the weaknesses found in Classical Test Theory (CTT). In IRT, the difficulty level and the discrimination power of test items remain consistent, even when the items are administered to different groups of test takers.

According Embretson & Reise (2013), Item Response Theory was developed by George Rasch, based on the assumption that a student's success in answering test items

depends on their ability and the difficulty level of the questions. Students with higher ability are expected to answer more questions correctly compared to those with lower ability.

Item characteristic analysis using IRT also measures the difficulty level and discrimination power of the items, but the calculation method differs from that of Classical Test Theory. In IRT, the probability that a test taker with ability Θ answers an item correctly is denoted by P(Θ) and is related to item difficulty and discrimination. Items that consider only the difficulty parameter are known as the Rasch model or 1-parameter model, and are formulated as follows:

$$P(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}, \quad \text{for } i = 1, 2, 3, ..., n$$

Explanation:

P(θ): probability that a test taker with ability level θ answers item i correctly
θ: ability level of the test taker
bi : difficulty index of item i
e: the natural number approximately equal to 2.718
n: number of items in the test

METHOD

This study is classified as an evaluative research, with the components being evaluated consisting of the students and the quality of the test items. The evaluation theory employed is the Rasch Model, selected due to its capability to evaluate both item quality and student ability, as well as its advantage in addressing limitations found in classical test theory.

The subjects of this study were 12th-grade students at Madrasah Aliyah Negeri 1 Surakarta, based on their responses to final exam questions in the Tarikh (Islamic history) subject. The sample consisted of 332 respondents, and according to Linacre (1994), this number meets the minimum sample requirement for Rasch Model analysis, which is 250 students.

The research variable is the item difficulty level. Data was collected using documentation techniques, specifically by gathering student responses to a final exam that had been developed by the teachers of Madrasah Aliyah Negeri 1 Surakarta.

Data analysis was conducted using the Rasch Model approach with the assistance of Winstep software. The aspects, measurement tools, and evaluation criteria used in this study. Validity and Reliability as Core Evaluation Aspects. These are the two main aspects that must be thoroughly analyzed to assess the quality of the evaluation instrument. Instrument validity concerns the accuracy of the items and the extent to which the test items align with the construct being measured. This was assessed using the Item Fit Order, which refers to three indicators: Outfit MNSQ: Ideally between 0.5–1.5, with values closer to 1 considered excellent, Outfit ZSTD: Ideally in the range of -2.0 to +2.0, Point Measure Correlation (Pt Mean Corr): Considered good if it falls

between 0.4 and 0.85. Unidimensionality refers to whether all items measure a single latent trait or ability. This is evaluated through, Raw variance explained by measures: Ranges from weak (<50%) to perfect (>80%). Unexplained variance: Considered weak if above 15%, and perfect if below 3% (Muntazhimah, Putri, & Khusna, 2020; Aprilia, 2021). Item Validity, Two key indicators are used here, Item bias, evaluated using Differential Item Functioning (DIF) and Differential Person Functioning (DPF). Items are considered biased or unfair if their probability value is greater than 0.05. Item accuracy, once again assessed through the same item fit indicators as above (Fernanda & Hidayah, 2020; Kirom, 2021). Item difficulty is measured using the Item Measure and categorized as follows: Very easy: Below -SD, Easy: Between -SD and mean, Moderate: Around the mean, Difficult: Between mean and +SD, Very difficult: Above +SD (Muntazhimah et al., 2020; Sabekti, 2018)

Reliability includes both, Person reliability (student response consistency), and

Item reliability (test consistency in differentiating abilities), with values categorized as follows: Weak: < 0.67, Good: 0.67–0.80, Very good: 0.81–0.94, Excellent: > 0.94. Additionally, overall test reliability is measured using Cronbach's Alpha, considered very good when above 0.80. Separation Index, This index measures the instrument's ability to distinguish between different levels of student ability, with an optimal value above 4.0, indicating a very good level of discrimination (Abdullah, Jahja, & Setiawan, 2022).

FINDINGS

Instrument Validity

Item validity can be observed from the alignment between respondents' answer patterns and the Rasch model, as well as from the unidimensionality of the instrument. Based on the analysis using the Rasch Model (see Table 1), it was found that all item feasibility indicators fall within the acceptable range according to the predetermined criteria. The Infit MNSQ value of 0.99 indicates that the respondents' answer data aligns well with the model, as it falls within the ideal range of 0.5 to 1.5. Likewise, the Outfit MNSQ value of 1.01 also reflects a good fit, as it remains within the acceptable range of 0.7 to 1.3.

Furthermore, the Infit ZSTD value of 0 and the Outfit ZSTD value of -0.1 show no significant deviation in the data, as both values are within the tolerance range of -2.2 to 2.2. This indicates that the response patterns of the participants are sufficiently stable and consistent.

Overall, these results show that the analyzed test items are categorized as fit, both in terms of infit and outfit, and based on both MNSQ and ZSTD values. Therefore, the items can be considered valid and appropriate for use in measurement.

Tabel 1. Item Set Fit to the Model

Aspect	Criteria	Measurement Result	Description
Infit MNSQ	0,5-1,5	0,99	Fit
Outfit MNSQ	0,7-1,3	1,01	Fit
Infit ZSTD	-2,2 s/d 2,2	0	Fit

Oufit ZSTD	-2,2 s/d 2,2	-0,1	Fit

The measurement results from Winsteps as shown in the following diagram:

```
     SUMMARY OF 34 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     281.6     386.8         .00     .16       .99     .0   1.01    -.1 |
| S.D.      77.9        .4        1.42     .05       .13    1.9    .57    2.4 |
| MAX.     381.0     387.0        3.95     .42      1.47    7.4   3.84    7.9 |
| MIN.      45.0     386.0       -3.35     .12       .80   -2.6    .46   -2.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .17 TRUE SD    1.41  SEPARATION  8.41  Item   RELIABILITY  .99 |
|MODEL RMSE    .17 TRUE SD    1.41  SEPARATION  8.54  Item   RELIABILITY  .99 |
| S.E. OF Item MEAN = .25                                                     |
-------------------------------------------------------------------------------
```

Thus, all 34 multiple-choice items fit the Rasch model.

The analysis results of the "raw variances explained by measures" aspect show a value of 20.4%. Based on the criteria used, this value falls into the weak category, as it is below the minimum threshold of 50% which indicates an acceptable measurement quality, see Table 2.

In general, this value indicates that only a small portion of the total data variance can be explained by the measurement model. In the context of construct validity, this suggests that the model's ability to represent the data structure is still very limited. In other words, the majority of the variance in the data cannot yet be adequately explained by the measured construct or dimension.

The low percentage of explained variance indicates the need for further evaluation of the items and the overall instrument structure, such as by revising or adding items that are more representative of the measured construct.

Table 2 Unidimensionality

Aspect	Criteria	Measurement Result	Description
raw variances explained by measures	<50% (lemah) 50-60% (jelek) 60%-70% (bagus) 70%-80%(sangat bagus) >80%(sempurna)	20,4%	lemah

Item Validity

Item validity can be measured based on the item difficulty level, item accuracy/item fit with the model, and item bias. Based on the measurement results of the item difficulty level, the data were classified into four categories: Very Difficult, Difficult, Easy, and Very Easy. The analysis results show that most items fall into the Easy category, with 15 items or 44% of the total. This indicates that the majority of the items could be answered well by the respondents.

Meanwhile, 11 items or approximately 32% fall into the Difficult category, indicating that these items require a deeper understanding or specific skills to be answered correctly. Furthermore, 4 items (12%) each fall into the Very Difficult and Very Easy categories. The Very Difficult category reflects significant challenges in answering those items, while the Very Easy category suggests that the items are likely too simple for the respondents, see Table 3.

Overall, this distribution of measurement results illustrates that the item difficulty levels tend to be balanced, with a greater tendency toward the Easy category.

Table 3. Recapitulation of Item Difficulty Levels

Item Measure	Criteria	Measurement Result	Percentage
> +SD (1,42)	Very Difficult	4	12%
+SD(1,42) s/d Mean (0,00)	Difficult	11	32%
Mean (0,00) s/d -SD (-1,42)	Easy	15	44%
< -SD (-1,42)	Very Easy	4	12%

The classification results are as follows, see Table 4.

Table 4. Item Difficulty Levels

Item	Logit value	Criteria	Item	Logit value	Criteria
S17	3,95	Very Difficult	S33	-0,19	mudah
S8	2,55	Very Difficult	S7	-0,29	Easy
S39	2,55	Very Difficult	S24	-0,29	Easy
S25	2,25	Very Difficult	S22	-0,31	Easy
S11	1,34	Difficult	S13	-0,33	Easy
S26	1	Difficult	S18	-0,41	Easy
S31	0,88	Difficult	S29	-0,61	Easy
S32	0,88	Difficult	S28	-0,68	Easy
S40	0,77	Difficult	S10	-0,94	Easy
S15	0,61	Difficult	S19	-1,03	Easy
S36	0,38	Difficult	S9	-1,32	Easy
S23	0,27	Difficult	S16	-1,36	Easy
S38	0,2	Difficult	S20	-1,36	Easy
S37	0,17	Difficult	S12	-1,47	Very easy
S34	0,13	Difficult	S21	-1,82	Very easy
S35	-0,06	Easy	S14	-1,97	Very easy
S30	-0,13	Easy	S27	-3,35	Very easy

There are two outlier items, namely item S17 and item S27. These items do not distinguish individuals effectively, as they have low measurement ability. The criteria for identifying outlier items and the results of the identification are as follows, see Table 5.

Table 5. Outlier Items

Scor	Criteria	Item
>2 SD (2,84)	> 2,84	S17

< -2 SD (-2,84)	< -2,84	S27

Based on the item validity analysis using the Rasch Model approach, the quality of the test items is described through three main indicators: Outfit Mean Square (MNSQ), Outfit Z-Standardized (ZSTD), and Point Measure Correlation (PT Corr). The criteria for item fit are as follows: Outfit MNSQ is within the range of 0.5 – 1.5 (ideally 1), Outfit ZSTD is within the range of -2 to +2 (ideally 0), and PT Corr is greater than 0.4 and not negative. Out of the 34 items analyzed, 32 items were found to fit the Rasch model. This means that these items fall within the tolerance range of the three established criteria, or at least only show minor violations in one aspect while still having a positive correlation with the measurement construct. However, there are 2 items identified as misfits, namely item S17 and item S39:

Item S17 has an Outfit MNSQ = 3.84, ZSTD = 7.2, and PT Corr = -0.04, indicating that this item deviates significantly from the ideal model and even has a negative correlation with ability scores, making it highly unsuitable for the intended measurement. Item S39 shows an Outfit MNSQ = 2.10, ZSTD = 7.9, and PT Corr = 0.00, indicating a high level of misfit. Although the correlation is not negative, it is zero, suggesting that this item does not contribute to improving the ability estimation. These two items should be considered for revision or removal from the instrument, given their deviation from the model and potential to compromise overall measurement accuracy. In general, the quality of the instrument is considered good, as the majority of items fall into the fit category, meaning they function well in measuring the intended construct, see Table 6.

Table 6. Misfit Items

Item	Outfit MNSQ Criteria: 0,5-1,5	Outfit ZSTD Criteria: -2 – 2	PT Correl Criteria >0,4	Description
S17	3.84	7.2	-.04	misfit
S39	2.10	7.9	.00	misfit
S8	1.39	3.3	.32	fit
S25	1.34	3.4	.37	fit
S32	1.17	2.1	.37	fit
S30	1.11	.8	.33	fit
S37	1.08	.7	.36	fit
S15	1.03	.4	.39	fit
S26	1.07	1.0	.40	fit
S35	1.02	.2	.37	fit
S28	1.06	.4	.37	fit
S9	.99	.1	.29	fit
S7	.90	-.6	.42	fit
S10	.87	-.5	.36	fit
S33	1.00	.0	.41	fit
S20	.81	-.6	.34	fit
S40	.93	-.8	.48	fit
S21	.87	-.3	.30	fit

S24	.87	-.8	.44	fit
S23	.82	-1.7	.50	fit
S27	.56	-.7	.21	fit
S12	.74	-.9	.37	fit
S11	.91	-1.3	.53	fit
S34	.79	-1.8	.51	fit
S19	.77	-1.0	.42	fit
S16	.68	-1.2	.40	fit
S31	.85	-1.9	.55	fit
S14	.46	-1.7	.37	fit
S18	.74	-1.7	.50	fit
S36	.74	-2.7	.55	fit
S29	.63	-2.3	.50	fit
S13	.70	-2.0	.52	fit
S38	.82	-1.6	.55	fit
S22	.62	-2.8	.56	fit

The misfit items are illustrated in the diagram below, see figure 1.

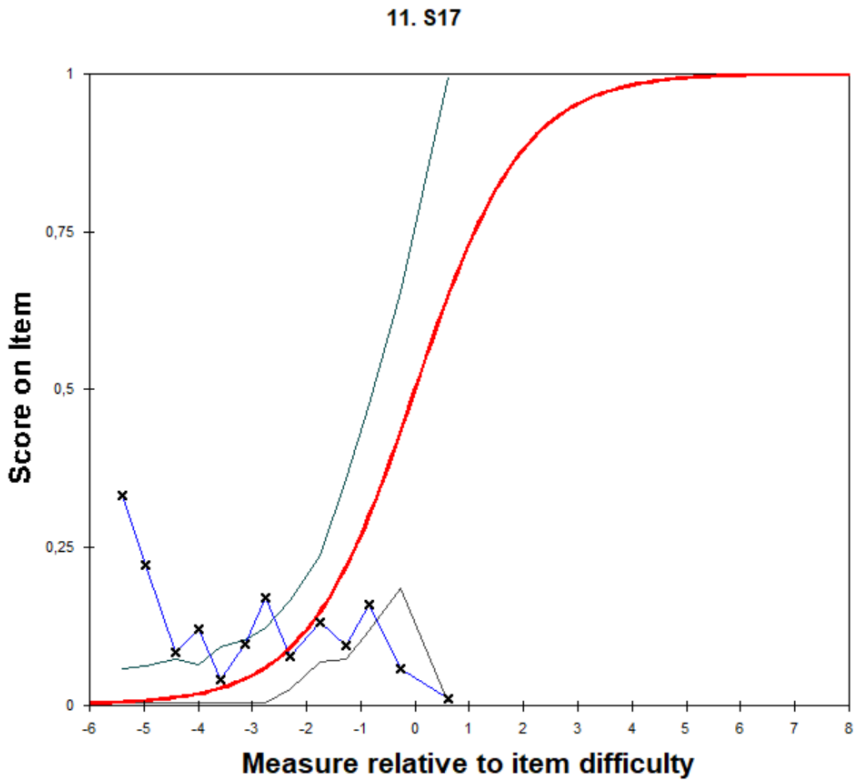

Figure 1. Biased Item of 17

Differential Item Functioning (DIF) analysis was conducted to examine whether a particular item treats different respondent groups fairly—for example, based on gender, educational background, or other categories. The criterion used in this analysis is a probability value (p-value) of < 0.05, which indicates that the item contains DIF or is statistically biased. From the analysis of 40 items, it was found that 19 items showed indications of containing DIF, meaning they had a p-value < 0.05. These items are: S7, S9, S11, S12, S13, S17, S19, S21, S25, S26, S28, S30, S31, S32, S33, S34, S35, S36, and S39.

The presence of DIF in these items indicates potential unfairness in measurement, where the item may favor or disadvantage certain groups, even if they possess the same level of ability. Meanwhile, the remaining 21 items were found to be free of DIF (NON DIF), as they had p-values above 0.05. These items are assumed to contribute to fair and equal measurement for all respondents, namely: S8, S10, S14, S15, S16, S18, S20, S22, S23, S24, S27, S29, S37, S38, and S40. It is noteworthy that item S27 had a p-value of 10, which is technically far above the threshold and clearly categorized as NON DIF, although the value appears unusual and may require further review for its validity, see Table 7.

Thus, based on the overall analysis, it can be concluded that more than half of the items are free from bias, but nearly half require further attention—either through revision or elimination—to ensure fairness and the validity of the measurement instrument.

Table 7. Biased Items (DIF)

Item	Prop.	Description	Item	Prop.	Description
S7	0,0026	DIF	S24	0,4784	Non DIF
S8	0,6767	Non DIF	S25	0	DIF
S9	0,0129	DIF	S26	0	DIF
S10	0,2387	Non DIF	S27	10	Non DIF
S11	0,0429	DIF	S28	0,0008	DIF
S12	0,0426	DIF	S29	0,0986	Non DIF
S13	0,0008	DIF	S30	0,035	DIF
S14	0,6393	Non DIF	S31	0,0127	DIF
S15	0,0724	Non DIF	S32	0,0002	DIF
S16	0,3278	Non DIF	S33	0,0094	DIF
S17	0,0363	DIF	S34	0,0188	DIF
S18	0,9785	Non DIF	S35	0,0342	DIF
S19	0,0406	DIF	S36	0,0036	DIF
S20	0,0601	Non DIF	S37	0,4745	Non DIF
S21	0,0398	DIF	S38	0,218	Non DIF
S22	0,3505	Non DIF	S39	0	DIF
S23	0,834	Non DIF	S40	0,108	Non DIF

The biased items are also illustrated in the graph below. The presence of non-identical curves indicates the existence of bias, see Figure 2.

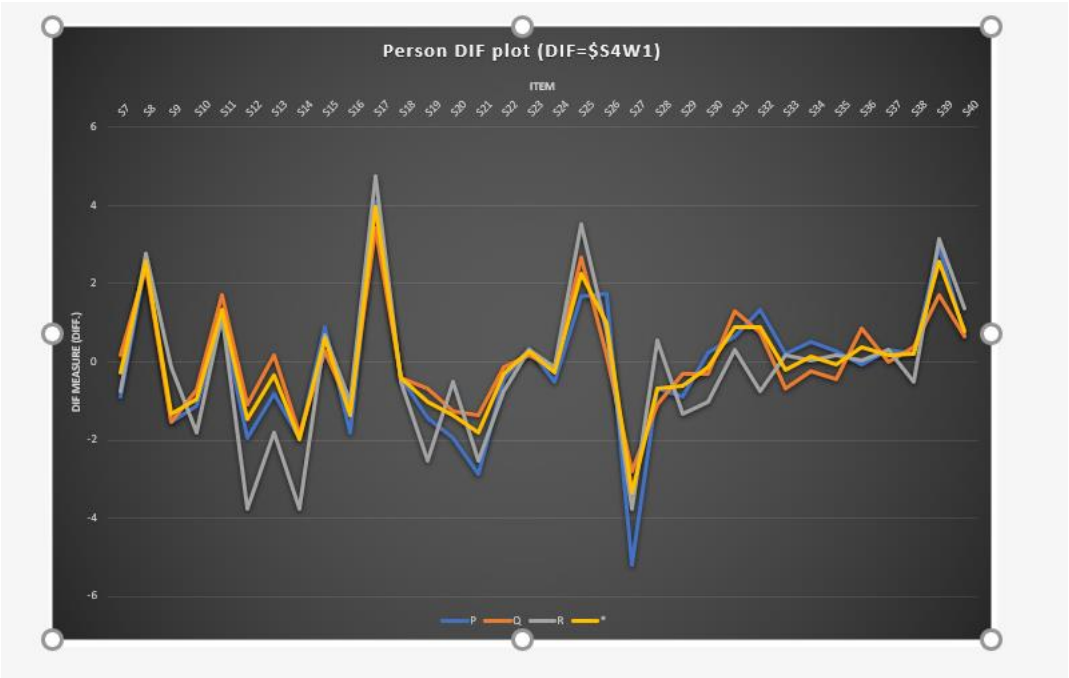

Figure 2. Biased Items

The summary is as follows, see Table 7.

Table 7. Recapitulation of Invalid Items

No	Aspect Analyzed	Unfit Item	Desciption
1	Tingkat kesulitan butir	S17, S27	Outlier
2	Tingkat keakuratan/ketepatan butir	S17, S39	misfit
3	Deteksi butir bias	S7, S9, S11, S12, S13, S17, S19, S21, S25, S26, S28, S30, S31, S32, S33, S34, S35, S36, dan S39.	bias

Reliability

The reliability analysis results indicate that the instrument used has very good quality, viewed from three main aspects: person reliability, item reliability, and Cronbach's Alpha. The person reliability value is 0.79, which falls within the range of 0.67–0.80, categorized as moderate. This indicates that the consistency of respondents' answers to the test items is at an adequate level. Although not very high, this value suggests that the respondents' abilities are reasonably well distributed within the instrument. The item reliability value reaches 0.99, which is in the excellent category (value > 0.94). This shows that the items in the instrument are highly reliable in distinguishing respondents with various ability levels. The test items demonstrate very high consistency and optimally present a range of difficulty levels. The Cronbach's Alpha value is 0.89, classified as very good (value > 0.8). This means that the internal consistency among items in the instrument is very strong, and overall, the instrument is stable and reliable in measuring the intended construct.

Therefore, overall, this measurement instrument is considered to have high reliability and is suitable for use, both from the perspective of respondents' responses and the quality of the test items themselves.

```
TABLE 3.1 D:\penelitian IIM 2023\AQIDAH DATA RAS ZOU700WS.TXT  Jul 31 20:38 2023
INPUT: 387 Person  40 Item  REPORTED: 387 Person  34 Item  2 CATS  WINSTEPS 3.73
-------------------------------------------------------------------------------

     SUMMARY OF 387 MEASURED Person
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      24.7      34.0        1.47     .51       .96     .0   1.01     .2 |
| S.D.       5.4        .4        1.17     .11       .25     .9    .88    1.0 |
| MAX.      33.0      34.0        4.55    1.09      1.73    2.9   9.90    5.3 |
| MIN.       8.0      27.0       -1.58     .39       .45   -1.8    .07   -1.4 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .54 TRUE SD    1.04  SEPARATION  1.93  Person RELIABILITY  .79 |
|MODEL RMSE    .52 TRUE SD    1.04  SEPARATION  2.01  Person RELIABILITY  .80 |
| S.E. OF Person MEAN = .06                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .84

     SUMMARY OF 34 MEASURED Item
-------------------------------------------------------------------------------
|          TOTAL                          MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     281.6     386.8         .00     .16       .99     .0   1.01    -.1 |
| S.D.      77.9        .4        1.42     .05       .13    1.9    .57    2.4 |
| MAX.     381.0     387.0        3.95     .42      1.47    7.4   3.84    7.9 |
| MIN.      45.0     386.0       -3.35     .12       .80   -2.6    .46   -2.8 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .17 TRUE SD    1.41  SEPARATION  8.41  Item   RELIABILITY  .99 |
|MODEL RMSE    .17 TRUE SD    1.41  SEPARATION  8.54  Item   RELIABILITY  .99 |
| S.E. OF Item MEAN = .25                                                     |
-------------------------------------------------------------------------------
```

Separation

The analysis results of person separation and item separation provide an overview of the instrument's quality in differentiating respondents' abilities and distinguishing item difficulty levels. The person separation value is 3.0, which falls into the moderate category according to the criteria of values between 2.0 and 3.0. This indicates that the instrument has adequate ability to differentiate groups of respondents based on their ability levels. Although it has not reached a high category, this value is sufficient to identify at least three different ability groups among the participants. The item separation value is 12.0, categorized as very high and even exceptional (value > 5.0). This indicates that the instrument has excellent precision in identifying various levels of item difficulty. In other words, the instrument can map items from the easiest to the most difficult very clearly and consistently. Thus, it can be concluded that this instrument is very strong in differentiating item characteristics (difficulty levels) and sufficiently reliable in distinguishing participant abilities, making it suitable for measurements requiring accurate classification, see Table 8.

Table 8. Item Separation Summary

separation	Criteria	Measurement Result	Criteria

Person separation Number of groups: ((4xseparasi)+1)/3	N<2(poor) 2,0-3,0 (fair) 3,0-4,0 (good) 4,0-5,0(very good) >5.0 (excellent)	=((4x1,93)+1)/3 =2,9 =3	fair
Item Separation Number of groups: ((4xseparasi)+1)/3	N<2(poor) 2,0-3,0 (fair) 3,0-4,0 (good) 4,0-5,0(very good) >5.0 (excellent)	=((4X8,41)+1)/3 =11,5 =12	excellent

DISCUSSION

Deviation of Items from the General Pattern (Outliers)

In the Item Response Theory (IRT)-based item analysis, questions number 17 and 27 on the Islamic history exam showed several indications of being problematic items, categorized as outliers. The questions are as follows:

17. Mundurnya Daulah Usmani ditandai dengan kebangkitan bangsa Barat atau Eropa (Renaisance), Runtuhnya Daulah Usmani disebabkan faktor faktor berikut kecuali ...

A Kondisi dan sistem pemerintahan yang lemah
B Kemerosotan dan lemahnya para pemimpin Daulah Usmani
C Semangat persatuan dan kesatuan seluruh pemimpin Daulah Mughal
D Melemahnya kekuatan militer dan serangan bangsa Eropa
E Munculnya gerakan oposisi Sekuler.

This question is classified as an outlier because its response pattern statistically deviates from other questions. Only 2% of students chose the correct answer (option C), while 63% selected A (an incorrect answer), see Table 9. When the majority of students—including those with high ability—answer incorrectly, despite the expectation that higher ability should correlate with a higher probability of correct answers, the item does not align with the IRT predictive distribution.

Table 9. Student response distribution for question 17

Question No. 17	Option	Number of Respondents	Percentage
	A	210	63%
	B	33	10%
	C	5	2%
	D	61	18%
	E	23	7%

IRT predicts that students with higher ability (high θ) will more frequently answer correctly than students with lower ability. However, this item shows the opposite pattern. Many high-ability students answered incorrectly because they were misled by the wording or focused too much on substantive content without noticing irregularities in the options. Conversely, some low-ability students answered correctly simply because they noticed the prominent but irrelevant name “Mughal.” This causes

the item characteristic curve (ICC) for this question to deviate and not match the general pattern of other items.

The discrimination parameter (a) for this item could be very low or even negative, as the item fails to differentiate between high- and low-ability students. The difficulty parameter (b) becomes inaccurate because the difficulty is not based on substantive content but on semantic traps. The guessing parameter (c) may be high, indicating students can guess correctly by detecting oddities.

Question number 27 is also identified as an outlier. The item reads:

27. *Pernyataan berikut yang menunjukkan bahwa generasi muda saat ini perlu meneladani para wali sanga dalam menyebarkan agama Islam adalah ...*
    - *A Kecerdasan dan kecerdikan para wali sanga dalam menyampaikan dakwah sehingga dapat diterima oleh masyarakat*
    - *B Ketegasan para wali sanga dalam berdakwah sehingga Islam tidak boleh kompromi dengan adat istiadat Jawa*
    - *C Para wali sanga banyak menguasai ilmu magic*
    - *D Para wali songo selalu mengikuti apa yang nenjadi kemauan masyarakat*
    - *E Semua wali sanga membekali dirinya dengan ilmu olah kanuragan sehingga ditakuti masyarakat*

In IRT-based evaluation, a question is called an outlier if the pattern of student responses significantly deviates from the model's prediction, which requires correct answers to correlate with student ability.

By construction, option A is clearly the correct answer as it is the only statement that is positive, rational, and aligns with Islamic da'wah values. In contrast, options B through E contain negative, irrational, or unscientific elements such as "magic," "feared by the people," or "no compromise." This imbalance makes the correct answer too obvious, so students may guess correctly without truly understanding the essence of the Wali Sanga's da'wah.

This situation means the question no longer measures substantive student ability but only tests the skill of identifying the most logical option. As a result, both high- and low-ability students have an equal chance of answering correctly, which flattens the item characteristic curve (ICC). The discrimination parameter (a) is very low, approaching zero, because the item fails to distinguish students by ability.

This question also violates the IRT unidimensionality assumption because the probability of a correct answer depends on basic logic or intuition, not mastery of the history or religious competence that should be measured. Furthermore, the highly uniform response pattern (78% of students answered E) causes the item to lose informative value in the IRT measurement system, see Table 10.

Table 10. Student response distribution for question 27

Question No. 27	Option	Number of Respondents	Percentage
	A	1	0%
	B	67	20%
	C	4	1%
	D	0	0%

E	260	78%

This strongly aligns with previous research findings emphasizing the importance of balanced options and item construction in IRT models. Studies by Hambleton et al. (1991) and Sumintono & Widhiarso (2014) highlight that items with overly obvious or disproportionate options may cause item misfit—where empirical data does not match IRT model predictions. A flat ICC and very low discrimination parameter indicate the item cannot distinguish students by ability. This also violates unidimensionality since the correct answer depends on logical intuition, not substantive competence in Islamic da'wah history. Therefore, these findings reinforce the recommendation that such items should be revised or removed to maintain the validity and reliability of the assessment instrument. (Embretson & Reise, 2013) explain in Item Response Theory for Psychologists that items with low discrimination or flat ICC typically lack adequate diagnostic value because they fail to differentiate between high- and low-ability students. Such items degrade measurement quality and should not be used. (Cai, Choi, Hansen, & Harrell, 2016) emphasize the need for caution in designing options so they are not too obvious or predictable. If students can guess the correct answer based on general logic or wording style, the item no longer measures the intended competence. (Holland & Wainer, 2012), in their study on Differential Item Functioning (DIF), show that items relying on specific experiences or backgrounds can create unfair chances of correct answers among groups with equal academic ability. This is also evident in the items you analyzed.

Because questions 17 and 27 are outliers, they need to be revised. The suggested item revisions are as follows. Revised item for question 17:

Berikut ini merupakan faktor-faktor utama yang menyebabkan runtuhnya Daulah Usmani, kecuali...
A. Melemahnya sistem pemerintahan
B. Kepemimpinan yang korup dan tidak visioner
C. Serangan militer dari kekuatan Eropa
D. Persatuan para pemimpin Daulah Mughal
E. Munculnya ideologi sekularisme dalam masyarakat

Revision note: Avoid the "except" construction (positive affirmation). Option D remains the correct answer (since it is irrelevant to the Ottoman Empire) but now the wording is more consistent, and the Mughal context remains as a clearer distractor without obscuring the main concept. To improve further, removing the Mughal option could be considered, since not all students may be familiar with Indian history.

For question 27, the options should be balanced and logical so that the correct answer cannot be easily guessed but requires understanding aligned with the intended competency. Suggested revision, Revised item for question 27:

Pernyataan berikut yang paling mencerminkan keteladanan para Wali Sanga dalam berdakwah yang relevan untuk diteladani generasi muda saat ini adalah ...
A. Menggunakan pendekatan budaya dan seni dalam menyampaikan ajaran Islam secara damai
B. Menolak semua bentuk tradisi lokal karena tidak sesuai dengan syariat
C. Berdakwah dengan kekuatan fisik agar ditakuti oleh masyarakat
D. Mewajibkan masyarakat menerima Islam secara keras dan mutlak

E. Menghindari dialog dan hanya fokus pada ceramah satu arah

The advantage of this revision is that distractors (B–E) remain incorrect but are more serious and realistic in language and content—without "magic" or "martial arts," which made the original item feel unacademic. Option A remains correct and Islamic in value but not too obvious—students must think and weigh their answer.

Mismatch between response patterns and ability (Misfit)

Question number 39 is indicated as misfit. The item reads:

39. Kaum Muslim di Amerika Serikat dan Kanada berhasil membentuk suatu wadah yang dapat memayungi seluruh kegiatan mereka. Kelompok ini dibuat oleh para imigran dan bermarkas di kota Plainfied Indiana. Organisasi paling besar ini bernama ...

A Islamic Society of North America (ISNA)
B American Society of Muslims (ASM)
C Islamic Circle of North America (ICNA)
D Islamic Assembly of North America (IANA)
E Muslim Public Affair Council (MPAC)

This question is classified as misfit because student responses do not match IRT model predictions. IRT assumes that higher-ability students have a higher probability of answering correctly. However, this item requires memorization of the organization's name, acronym, and headquarters location (ISNA). High-ability students with strong understanding of Islamic history or global da'wah may answer incorrectly if they do not memorize these details. Conversely, lower-ability students who have heard the name "ISNA" may guess correctly. This pattern lowers item discrimination and causes the ICC to deviate, resulting in statistical misfit.

In the IRT context, as explained by Embretson & Reise (2013), items relying too much on factual memorization tend to have low discrimination and distorted ICC because they fail to reflect the substantive ability intended to be measured—in this case, understanding of Islamic history or da'wah insight. Cai et al. (2016) also show such items can cause misfit by measuring a dimension not aligned with measurement goals.

Therefore, this item needs revision to measure conceptual understanding rather than specific, contextual factual memorization. Suggested revision for question 39:

Salah satu bentuk perkembangan Islam di negara-negara minoritas Muslim seperti Amerika Serikat adalah berdirinya organisasi Islam yang mewadahi kegiatan sosial dan dakwah. Organisasi Islam terbesar di kawasan Amerika Utara adalah...
A. Islamic Society of North America (ISNA)
B. Majelis Ulama Amerika (MUA)
C. Dewan Syariah Internasional (DSI)
D. Muslim Brotherhood of America (MBA)
E. Islamic Forum of North America (IFNA)

This revised question focuses on the general concept of global Islamic organizations, emphasizing organizational function and region—not headquarters or founders.

Bias

In an analysis of several exam items (no. 7, 9, 11, 12, 13, 17, 19, 21, 25, 26, 28, 30, 31, 32, 33, 34, 35, and 36), various forms of bias were identified that have the potential to undermine the fairness of measuring students' abilities. These biases can be grouped into several categories:

1. Inequitable Access to Information

Item number 17 is biased because it contains a "trap" option and requires specific historical knowledge without sufficient context, making it unfair for students from different backgrounds. Item number 39 is also biased due to cultural/geographical context bias. It tests knowledge of an Islamic organization in North America, which is likely unfamiliar to most students in Indonesia unless they have specifically studied global Muslim issues. This disadvantages students who lack access to global information, even if they possess strong religious knowledge.

Furthermore, it is irrelevant to the local curriculum context. If this question appears in the context of Islamic religious education or Islamic history in Indonesia, the content is disproportionate to the local students' context. This leads to assessment unfairness.

2. Dependence on Specialized Knowledge and Memorization of Historical Facts

Many items (such as no. 7, 9, 11, 13, 25, 26, 28, 30, 35, and 36) require students to memorize names of historical figures, treaties, or titles in Islamic and Indonesian history. For example, item no. 28 asks for the birth name of Sunan Muria without context, and item no. 26 requires the memorization of a long title of a Demak Sultan. Students unfamiliar with such historical details may struggle to answer correctly.

3. Similar or Confusing Answer Choices

Items no. 7 and 11 present ambiguous answer choices, such as spiritual facts being considered economic evidence, or treaty names that sound similar. Items no. 12 and 21 also include vague terms like "Revival Movement" without clarification. This increases the risk of misinterpretation and inaccurate responses.

4. Complex Terminology Without Contextual Explanation

Some items (no. 7, 9, 11, 17, 21) use technical terms such as "Ukaz Market," "false prophets," or "Propaganda Movement" without providing definitions or context, making it difficult for students who haven't formally learned these terms. The absence of cues gives an unfair advantage to those who are already familiar.

5. Unequal Cultural or Educational Backgrounds

Items such as no. 19 (about Akbar's tolerance in India), no. 31–34 (questions about Malaysia, Singapore, Myanmar, and Korea), and no. 35 (about Abyssinia) test highly specific knowledge that may not be taught equally across all schools or regions. This leads to Differential Item Functioning (DIF), where students from different groups have varying chances of answering correctly, even with the same ability level.

6. Sensitive and Specific Religious Topics

Item no. 9 involves the theme of "false prophets," which is religiously sensitive. Without proper context or careful handling, such questions may appear biased and non-inclusive to students from different backgrounds.

7. Lack of Context for Time, Place, or Narrative

Items like no. 25 and 26, which discuss the history of Islamization and royal titles, as well as item no. 21 on the Islamic revival movement, often lack contextual cues about time or region, making it difficult for students to connect the information.

This issue aligns with findings and critiques in previous studies on construct validity and fairness in the development of history or religious education items based on Item Response Theory (IRT). In the literature, Haladyna, Downing, & Rodriguez (2002) emphasize the importance of fair, contextualized, and culturally unbiased item construction. Overreliance on memorization of facts and names, as seen in items no. 28 and 26, can reduce item discrimination and violate the principle of unidimensionality, as criticized by Embretson & Reise (2013). The use of complex terms without context, ambiguous answer choices, and references to foreign regions or figures not taught uniformly are common causes of Differential Item Functioning (DIF), as discussed by Holland & Wainer (2012) and Van de Vijver & Tanzer (2004). This means that students with certain backgrounds (e.g., strong formal Islamic education or international exposure) will have an advantage that does not reflect the actual ability being measured.

In addition, sensitive topics like “false prophets” without adequate contextual framing can violate the principles of inclusivity and item neutrality, as warned by Subali (2010) in his study on test item bias and Yulianto (2020) in their research on bias in psychological scales. In conclusion, these items are biased because they tend to assess memorization of specific facts or recognition of niche terms rather than fairly evaluating conceptual understanding. Similar answer choices, lack of contextual support, and unequal background experiences among students render these items universally invalid. To ensure fair evaluation, exam items must be designed with clarity, context, and awareness of the diverse backgrounds of test-takers.

Test Package Reliability

Meanwhile, the results of the reliability analysis indicate that the instrument used has very good quality based on three main indicators: person reliability, item reliability, and Cronbach’s Alpha. The person reliability score of 0.79 falls within the “adequate” category, indicating that participants’ responses were reasonably consistent. Although not exceptionally high, this value suggests that the instrument successfully captures the range of participants' abilities. Next, the item reliability score of 0.99 is considered excellent. This shows that the test items are highly effective in distinguishing participants based on different levels of ability. The high score also reflects strong consistency and high quality in the items themselves. The Cronbach’s Alpha value of 0.89 indicates a very strong level of internal consistency among the items. This means that the instrument as a whole is stable and reliable in measuring the intended construct.

These findings are consistent with the views of Sumintono & Widhiarso (2014), who stated that high reliability—at both the individual and item levels—is a key requirement in developing instruments based on the Rasch model. This is further supported by Azwar (2016) and Budiastuti (2022), who emphasize the importance of internal consistency in ensuring the reliability of measurement results. Based on these

results, it is recommended that future studies strengthen construct validity and revise the test items to avoid bias in order to ensure fairness across different respondent groups. In addition, developing an item bank and testing the instrument in varied contexts is also important for enhancing generalizability. Thus, this instrument is not only statistically reliable, but also relevant and fair in practice.

CONCLUSION

This study aims to analyze the quality of Fiqh test items for Madrasah Aliyah students, consisting of 40 multiple-choice questions. The analysis was conducted based on three main aspects: difficulty level, item fit, and bias detection. The results show that two items (S17 and S27) are outliers in terms of difficulty level, two items (S17 and S39) are misfits or do not conform to the general response pattern of the students, and 19 other items are indicated to contain bias, which may affect the fairness of the measurement. Regarding reliability, the instrument demonstrated excellent results. A person reliability of 0.79 indicates that student responses were fairly consistent. A high item reliability of 0.99 suggests that the test items are very effective in distinguishing students based on their ability levels. Furthermore, a Cronbach's Alpha value of 0.89 reinforces that the instrument has very good internal consistency. The study recommends revising the items categorized as misfit, outliers, and biased, particularly items S17 and S39, which appeared in more than one category of discrepancy.

REFERENCES

Abdullah, N., Jahja, M., & Setiawan, D. G. E. (2022). Analisis kualitas butir soal pada mata pelajaran fisika di jurusan fisika fakultas mipa universitas negeri gorontalo tahun ajaran 2021/2022. *Jurnal Sains Dan Pendidikan Fisika, 18*(1), 44-52. doi: https://core.ac.uk/download/pdf/522888893.pdf

Alfarisa, F., & Purnama, D. N. (2019). Analisis butir soal ulangan akhir semester mata pelajaran ekonomi SMA menggunakan Rasch model. *Jurnal Pendidikan Ekonomi Undiksha, 11*(2), 366-374. doi: http://dx.doi.org/10.23887/jjpe.v11i2.20878

Azwar, S. (2016). Reliabilitas dan validitas aitem. *Buletin Psikologi, 3*(1), 19-26.

Budiastuti, D. (2022). *Validitas dan reliabilitas penelitian*.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application, 3*(1), 297-321.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*: Psychology Press.

Fernanda, J. W., & Hidayah, N. (2020). Analisis kualitas soal ujian statistika menggunakan classical test theory dan rasch model. *Square: Journal of Mathematics and Mathematics Education, 2*(1), 49-60. doi: https://doi.org/10.21580/square.2020.2.1.5363

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning:* Routledge.

Ibn Utsman Al-Faris, I. (2013). Murnikan Aqidah Luruskan Ibadah. Jakarta: Pustakan at-Tazkia.

Kurniaty, Y., & Praja, C. B. E. (2016). Keluarga Sebagai Agen Pembentuk Kader Muhammadiyah. *Jurnal Tarbiyatuna, 7*(1), 25-37.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch measurement transactions, 7*, 328.

Muntazhimah, M., Putri, S., & Khusna, H. (2020). Rasch model untuk memvalidasi instrumen resiliensi matematis mahasiswa calon guru matematika. *JKPM (Jurnal Kajian Pendidikan Matematika), 6*(1), 65-74. doi: https://journal.lppmunindra.ac.id/index.php/jkpm/article/view/8144

Novriyanti, E., & Arthur, R. (2024). Analisis kualitas butir soal ujian tengah semester Biologi umum menggunakan Model Rasch. *JagoMIPA: Jurnal Pendidikan Matematika dan IPA, 4*(4), 718-733. doi: https://doi.org/10.53299/jagomipa.v4i4.927

Parisu, C. Z. L., Ekadayanti, W., Sisi, L., & Juwairiyah, A. (2024). Analisis Butir Soal Pengetahuan Dasar Matematika Menggunakan Pendekatan Rasch. *Science Tech: Jurnal Ilmu Pengetahuan dan Teknologi, 10*(1), 36-45.

Rezky, F. N. D. (2025). Analisis Kebijakan tentang Ujian Nasional. *Jurnal Pendidikan, Sains, Geologi, dan Geofisika (GeoScienceEd Journal), 6(*2), 824-830.

Saepuzaman, D., Haryanto, H., Istiyono, E., Retnawati, H., & Yustiandi, Y. (2021). Analysis of items parameters on work and energy subtest using item response theory. *Jurnal Pendidikan MIPA, 22*(1), 1-9.

Subali, B. (2010). Bias item tes keterampilan proses sains pola divergen dan modifikasinya sebagai tes kreativitas. *Jurnal Penelitian dan Evaluasi Pendidikan, 14*(2), 309-334.

Subando, J., Saleh, M., & Ngatimin, N. (2022). The Analysis of Question Item on Tarikh Subject with Rasch Model Theory. *Journal of Research and Educational Research Evaluation, 11*(2), 166-177.

Sumintono, B., & Widhiarso, W. (2014). Model Rasch untuk penelitian sosial kuantitatif. Makalah Kuliah Umum Di Jurusan Statistika, ITS Surabaya, 21 November 2014, November, 201, 1-9.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European review of applied psychology, 54*(2), 119-135.

Widana, I. W. (2014). Pengembangan bank soal. *Emasains, 3*(2), 186-197.

Wisman, Y., Effrata, E., & Tutesa, T. (2021). Penerapan konsep instrumen evaluasi hasil belajar. *Jurnal Ilmiah Kanderang Tingang, 12*(1), 1-9.

Yulianto, A. (2020). Mewaspadai response bias dalam skala psikologi. *Buletin KPIN, 6*(03).

Yusuf, M. (2023). Evaluasi Metode Penilaian dalam Pendidikan Islam dalam Upaya Meningkatkan Ketepatan dan Objektivitas Penilaian Siswa. *Sasana: Jurnal Pendidikan Agama Islam, 2*(1), 77-82.