Proceedings of the 1st International Conference on Science and Islamic Studies
ISSN: 3024-9058 (media online)
Volume: 1, 2023
Publisher: Universitas Islam Negeri Alauddin Makassar

Comparison of item analysis of madrasah assessment questions between Classical Test Theory (CTT) and the Rasch Model (studies on akidah akhlak course)

Indriani H. Ismail[1], Muhammad Khalifah Mustami[2], Misykat Malik Ibrahim[3], Andi Achruh[4], Bahaking Rama[5], Yuspiani[6], Wahyuddin Naro[7], & Saprin Sagena[8]

[1]IAI DDI Pangkajene Sidenreng Rappang
[2,3,4,5,6,7,8]Universitas Islam Negeri Alauddin Makassar

Correspondence Email: indrianih.ismail0009@gmail.com

ABSTRACT

This study aimed to compare item analysis results between the CTT (Classical Test Theory) method and the Rasch model in assessing Akidah Akhlak course in Madrasahs. The study utilized data from the assessment questions of Akidah Akhlak course administered to 22 students. The questions consisted of 40 multiple-choice items with five answer options. The study employed a quantitative approach with a qualitative comparative design. The data were analyzed using the R Program. The research findings showed both differences and similarities between the CTT method and the Rasch model in analyzing the quality of the Akidah Akhlak course assessment questions. These differences and similarities could be observed in the parameters utilized by both approaches, such as difficulty index, discrimination index, distractor effectiveness, and item fit with the model. Generally, both approaches tended to agree in identifying questions as easy or difficult, discriminative or non-discriminative, and fitting or not fitting the model. However, some questions exhibited significant differences in parameter values between the CTT method and the Rasch model. This suggests that specific questions had different characteristics for test takers with varying abilities or different samples. This study contributes to developing and enhancing the quality of Akidah Akhlak assessment questions in Indonesia. Furthermore, it provides valuable information for question developers, teachers, students, and researchers regarding the strengths and weaknesses of the CTT method and the Rasch model in analyzing the quality of multiple-choice questions.

Keywords: Item analysis, CTT method; Rasch model; madrasah assessment; Islamic education

1. INTRODUCTION

Educational evaluation is a critical process to measure the quality and effectiveness of teaching and learning activities. Evaluation provides feedback and information to various stakeholders, such as teachers, students, parents, and policymakers, to improve educational outcomes and standards. The evaluation also ensures accountability and transparency of educational institutions and programs.

One form of educational evaluation in Indonesia is the Madrasah Assessment (Assessment Madrasah or AM), conducted by the Ministry of Religious Affairs for madrasah education levels. Madrasahs are Islamic schools that offer general education, Islamic religious education, and Arabic language courses. The AM aims to measure students' competence in *PAI* and Arabic courses, which are distinctive features of madrasah education. The AM also aims to improve the quality of madrasah education, provide feedback to stakeholders, and ensure the quality of madrasah graduates.

A critical aspect of the AM is the quality of the questions used. High-quality questions must meet the criteria of validity, reliability, difficulty level, discrimination power, and effectiveness of distractors. These criteria can affect the accuracy and fairness of test results and the validity and reliability of inferences made from the test results. To assess the quality of questions, item analysis is required, which can be performed using various approaches, including the classical test theory (CTT) and the Rasch model.

CTT (Classical Test Theory) is the most commonly used approach in item analysis. CTT assumes that the test score combines the valid score and measurement error. The accurate score reflects the actual ability or achievement of the test taker, while the measurement error represents random or systematic factors that affect the test score. CTT utilizes parameters such as item difficulty, discrimination power, and distractors' effectiveness to evaluate the questions' quality. Item difficulty indicates how easy or difficult a question is for test takers. Discrimination power measures how well a question can differentiate between high and low-ability test takers. The effectiveness of distractors indicates how well incorrect answer choices attract low-ability test takers.

The Rasch model is one of the models in Item Response Theory (IRT), which is an alternative to CTT. The Rasch model assumes that the probability of correctly answering an item is solely determined by the test taker's ability and the item's difficulty. Ability and difficulty are measured on a standard scale known as logits. The Rasch model utilizes parameters such as item fit to the model, which is assessed through infit MNSQ, outfit MNSQ, and Point measure correlation to evaluate the quality of the questions. Item fit to the model indicates how well an item conforms to the expectations of the Rasch model. Infit MNSQ and outfit MNSQ are statistics that measure the degree of deviation from model fit. Point Measure-correlation indicates how well an item correlates with the overall test score.

Researchers have conducted comparisons between CTT and the Rasch model in item analysis. Some studies have shown that the Rasch model can provide better analysis than

CTT in measuring reliability, difficulty level, and discrimination power. The Rasch model can offer more accurate and consistent estimations of these parameters across different samples and tests. The Rasch model can also detect misfitting or problematic items, which may require revision or removal. However, other studies have shown that CTT and the Rasch model produce similar or non-significantly different results. CTT and the Rasch model may agree in identifying items as easy or difficult, discriminative or non-discriminative, and effective or ineffective distractors.

This study aims to compare item analysis results between CTT and the Rasch model on the Madrasah's Assessment questions of the *PAI* subject group focusing on the Aqidah Akhlak subject. Both CTT and the Rasch model will be applied in this study for analysis. The study will compare the parameters obtained from both approaches and examine their similarities and differences. Based on both approaches, this study will also discuss the implications and recommendations for improving the quality of AM questions.

This research is expected to contribute to developing and improving the quality of Madrasah Examinations/Assessments for the *PAI* subject group. The study will provide empirical evidence regarding the strengths and limitations of CTT and the Rasch model in item analysis. Additionally, practical recommendations will be given for developing valid, reliable, fair, and effective items for both formative and summative assessments.

2. METHODS

This research adopts a quantitative approach with a qualitative comparative design. This design aims to compare item analysis results between CTT and the Rasch model on Madrasah Examinations/Assessments for the *PAI* subject group, focusing specifically on Aqidah Akhlak. The study utilizes a sample of AM questions from one of the madrasahs in Sidenreng Rappang regency. The total number of items is 45, and the respondents consist of 22 individuals.

The data collection procedure is as follows:

- The researcher obtained permission from the madrasah principal to use the AM questions and student answer sheets for this research.
- The researcher collects the AM questions and student answer sheets from the madrasahs and scans them into digital files.
- The researcher codes the questions and student answers using Microsoft Excel and analyzes the data using the R program.
- The researcher exports the data from Microsoft Excel to the R program in two formats: CTT analysis and Rasch model analysis.

The data analysis procedure is as follows:

- For CTT analysis, the researcher uses the R program with the CTT package to calculate the following parameters for each item: difficulty index and discrimination index. The difficulty index is calculated as the proportion of students who answered the item correctly. The discrimination index is calculated as the point biserial correlation between the item and total test scores.

- For Rasch model analysis, the researcher uses the R program with the ltm package to estimate the following parameters for each item: difficulty level and fit statistics. The difficulty level is measured on a logit scale, indicating how difficult or easy an item is relative to the average student's ability.
- The researcher compares each item's CTT and Rasch model analysis results and examines their similarities and differences. The researcher also identifies problematic or misfitting items according to one or both approaches and suggests ways to revise or remove them.

The ethical considerations of this research are as follows:

- The researcher ensures that the AM items and student answer sheets are used solely for research purposes, not other purposes.
- The researcher protects the confidentiality and anonymity of the madrasahs and students by not disclosing their names or identities in any report or publication.
- The researcher obtains written consent from the head of the madrasah and the students before using their data for this research.

The researcher adheres to the ethical guidelines of their institution and the Ministry of Religious Affairs in conducting this research.

3. RESULTS AND DISCUSSION

Table 1. Results of Item Analysis Using CTT

Item Number	Difficulty Index	Discrimination Index	Item Number	Difficulty Index	Discrimination Index
1	0,174	0,913	21,000	0,174	0,913
2	0,174	0,913	22,000	0,174	0,826
3	0,174	0,913	23,000	0,087	0,957
4	0,174	0,913	24,000	0,261	0,870
5	0,174	0,826	25,000	0,348	0,696
6	0,174	0,870	26,000	0,174	0,826
7	0,174	0,913	27,000	0,174	0,870
8	0,174	0,913	28,000	0,087	0,957
9	0,087	0,957	29,000	0,261	0,870
10	0,000	0,870	30,000	0,174	0,913
11	0,087	0,957	31,000	0,000	0,000
12	0,087	0,913	32,000	0,000	0,000
13	0,087	0,957	33,000	0,000	0,000
14	0,087	0,870	34,000	0,174	0,913
15	0,174	0,913	35,000	0,261	0,826

16	0,087	0,957	36,000	0,261	0,870
17	0,261	0,783	37,000	0,087	0,826
18	0,174	0,913	38,000	0,174	0,913
19	0,348	0,826	39,000	0,174	0,913
20	0,087	0,957	40,000	0,174	0,913

Table 2. Results of Analysis of Question Items with the Rasch Model

Item Number	Difficulty Index	Item Number	Difficulty Index	Item Number	Difficulty Index	Item Number	Difficulty Index
1	-2.673	11	-16.547	21	-2.673	31	20.235
2	-2.673	12	-2.673	22	-1.694	32	20.235
3	-2.673	13	-16.547	23	-16.547	33	20.235
4	-2.673	14	-2.087	24	-2.087	34	-2.673
5	-1.694	15	-16.547	25	-0.907	35	-1.694
6	-2.087	16	-16.547	26	-1.694	36	-2.087
7	-2.673	17	-1.388	27	-2.087	37	-1.694
8	-2.673	18	-2.673	28	-16.547	38	-2.673
9	-16.547	19	-1.694	29	-2.087	39	-2.673
10	-2.087	20	-16.547	30	-2.673	40	-2.673
Discrimination Index = 1.545							
Log.Lik: -151.772							

Based on the item analysis using CTT, the following information is obtained:

1. Difficulty level indicates the proportion of students who answered each item correctly. The higher the difficulty value, the easier the item is, and vice versa. Discrimination indicates the ability of the item to differentiate between high- and low-performing students. The higher the discrimination value, the better the item's discriminatory power, and vice versa. SD represents the standard deviation of scores for each item. Item total indicates the correlation between the item score and the total score. Item.Tot.woi indicates the correlation between the item score and the total score without that item.
2. The difficulty values range from 0 to 1, indicating that some items are straightforward, easy, complex, and challenging.
3. The discrimination values range from 0 to 0.571, indicating that there are items with very good, good, fair, and poor discriminatory power.
4. The items with a difficulty value of 1 are item numbers 9, 11, 13, 16, 20, 23, 28, and 40, indicating that these items are straightforward, and no students answered them incorrectly.
5. The items with a difficulty value of 0 are item numbers 31, 32, and 33, indicating that they are very difficult, and no students answered them correctly.

6. The items with the highest discrimination value are item numbers 17 and 19, with a value of 0.571, indicating that these items have excellent discriminatory power.

Based on the item analysis using the Rasch model, the following information is obtained:

1. The Difficulty coefficient indicates the difficulty level of each item, where higher values indicate more incredible difficulty and vice versa. The Discrimination coefficient indicates the ability of each item to differentiate between high- and low-performing students—the Log. Like (Log Likelihood) represents the maximum likelihood value of the Rasch model, where higher values indicate a better fit of the Rasch model.
2. The Discrimination value is 1.545, indicating that the items have a sufficiently good discriminatory power.
3. *The Log. Lik* value is -151.772, indicating that the Rasch model fits the data reasonably well.
4. The Difficulty values range from -16.547 to 20.235, indicating that some items are straightforward, easy, complex, and difficult.
5. The items with positive and significant Difficulty values are item numbers 31, 32, and 33, indicating that these items are very difficult.
6. The items with negative and small Difficulty values are item numbers 9, 11, 13, 16, 20, 23, 28, and 40, indicating that these items are very easy.

This study’s results indicate differences and similarities between the CTT method and the Rasch model in analyzing the quality of items in the Aqidah Akhlak subject of the AM examination. These differences and similarities can be observed in the parameters used by both approaches, specifically the difficulty index. The difficulty index reflects how easy or difficult an item is for test takers. In the CTT method, the difficulty index is calculated as the proportion of students who answered the item correctly. The difficulty index values range from 0 to 1, with higher values indicating easier items. In the Rasch model, the difficulty level is measured on a logit scale, indicating how difficult or easy an item is relative to the average ability of students. The difficulty level values range from minus infinity to plus infinity, with lower values indicating easier items.

This study indicates that both approaches tend to agree in identifying easy or difficult items. However, with more than one logit difference, some items show significant differences between the difficulty index in CTT and the difficulty level in Rasch. This suggests that some items exhibit different characteristics for test takers with different abilities.

Comparing these findings with a previous study by Nurhayati (2018) that also used CTT and Rasch model to analyze the quality of multiple-choice items, similar patterns can be observed. Nurhayati's study on analyzing mathematics items for high school students using CTT and Rasch model also found a strong positive correlation between the difficulty index in CTT and the difficulty level in Rasch, with a coefficient of 0.94. The study also identified several items with significant differences between the difficulty index in CTT and the difficulty level in Rasch.

These consistent findings across different studies highlight the importance of considering CTT and Rasch model analyses to understand item quality comprehensively. It indicates that while the two approaches generally agree, there may still be specific items that exhibit divergent characteristics based on the test takers' abilities.

The difference in the CTT difficulty index value and the Rasch difficulty level on several questions can be caused by several factors, such as:

- Item characteristics

Some items may have varying difficulty levels for test takers with different abilities. For example, items containing terms or concepts that are less familiar to test takers may be more challenging for low-ability test takers compared to high-ability test takers. This can result in higher values of the difficulty index in CTT compared to the Rasch difficulty level because CTT calculates the proportion of students who answered correctly without considering their abilities. In contrast, the Rasch model calculates the difficulty level relative to the average ability of students.

- discrimination index

The discrimination index indicates how well an item can differentiate between high-ability and low-ability test takers. In CTT, the discrimination index is calculated as the point-biserial correlation between the item and total test scores. The value of the discrimination index ranges from -1 to 1, with higher values indicating more discriminative items. In the Rasch model, the point-measure correlation is calculated as the correlation between the item score and the student's ability measure. The value of the point-measure correlation also ranges from -1 to 1, with higher values indicating more discriminative items.

To address the differences between CTT and the Rasch model in measuring the effectiveness of distractors, it is necessary to revise or develop items that take into account the characteristics of the items and the sample. For example, distractors or incorrect answer choices that are too far from the key answer can be replaced with distractors closer to the key answer but still incorrect. Distractors or answer choices that are irrelevant to the item content or contradict general knowledge can be replaced with distractors or answer choices that are relevant to the item content and aligned with general knowledge but still incorrect. Distractors or answer choices that are specific to a particular local or cultural context can be replaced with more general or neutral distractors or answer choices.

To address items not aligned with the Rasch model, it is necessary to revise or develop items that consider the characteristics of the items and the sample. For example, items with poor structure or format can be simplified or clarified to make them easier to understand and answer. Items related to topics that have not been learned or practiced beforehand can be adjusted to the sample's level of mastery, making them more relevant and fair.

CONCLUSIONS

Based on the above analysis, it can be concluded that the CTT method and the Rasch model have their respective strengths and limitations in analyzing the quality of multiple-choice items in the subject of Islamic Education (*PAI*), focusing on aqidah akhlak. The strengths of the CTT method are its ease of calculation and understanding and its applicability to various tests. The limitations of the CTT method are its dependence on specific samples and tests and its inability to measure ability and difficulty in an absolute sense.

On the other hand, the strengths of the Rasch model are its ability to measure ability and difficulty in an absolute and consistent manner and its capability to detect items that are not aligned with the model. The limitations of the Rasch model are its complexity in calculation and understanding and its reliance on certain assumptions to hold.

This research also provides several implications and recommendations for developing and improving the quality of multiple-choice questions in Islamic Education (*PAI*). Some of these implications and recommendations are as follows:

1. Questions need to be designed considering the characteristics of the test takers, who may come from different cultural backgrounds and contexts. In this regard, it is essential to avoid using terms, concepts, or contexts that are too specific or related to a particular culture that may not be familiar to all test takers. Questions should be formulated clearly and understood by all test takers.
2. The provided distractors or answer choices in the questions must be carefully revised. Distractors that are too easy or too far from the correct answer can reduce the discriminative power of the question. Distractors that are irrelevant or contradict common knowledge can confuse test takers. Therefore, it is essential to develop plausible yet incorrect distractors that can differentiate between test takers with good and poor understanding.
3. The process of analyzing the quality of questions using both the Classical Test Theory (CTT) and the Rasch model can provide more comprehensive information. Therefore, using both methods simultaneously in analyzing the quality of questions is recommended. This will help gain a better understanding of the difficulty and ability of questions and identify potential issues that may arise in measurement.
4. Further research can be conducted to deepen the understanding of the quality of multiple-choice questions in Islamic Education and test the effectiveness of using

CTT and the Rasch model in different contexts. More valid and reliable information about the quality of questions and the analysis methods used can be obtained by involving larger samples and broader variations in testing contexts.

5. Question developers and Islamic Education teachers need to maintain consistency in using methods to analyze the quality of questions and conduct regular evaluations of the questions used. By continuously evaluating and improving the questions, the quality can be enhanced to better measure test takers' understanding of Islamic Education courses.

It is important to note that these implications and recommendations are just some examples and can be adjusted to the research context and specific needs of question development.

REFERENCES

Amin, M., & Sutopo, S. (2019). Karakteristik Tes Kemampuan Berpikir Kritis Siswa SMA pada Materi Momentum dan Impuls Perbandingan Classical Theory Test (CTT) dan Model Rasch. Jurnal Pendidikan Fisika dan Teknologi, 5(1), 1-10.

Andriani, N. (2018). Analysis Butir Soal Ujian Akhir Semester Mata Kuliah Statistika Dasar Menggunakan Model Rasch (Doctoral dissertation, Universitas Gadjah Mada). http://etd.repository.ugm.ac.id/penelitian/detail/109065

Anwar, A. (2018). Analisis Butir Soal Ujian Akhir Semester Mata Kuliah Statistika Inferensial Menggunakan Model Rasch (Doctoral dissertation, Universitas Gadjah Mada). http://etd.repository.ugm.ac.id/penelitian/detail/111147

Arifin, Z., & Fauzi, A. (2020). Learning Assesment For Madrasah Strengthening Islamic Psycososial and Emotional Inteelligence. Journal of Education and Learning Studies, 3(2), 1-10. https://repository.uir.ac.id

Asrori, M., & Nurhayati, N. (2018). Penilaian Berbasis Kelas pada Pembelajaran Pendidikan Agama Islam di Madrasah. Jurnal Pendidikan Islam Indonesia, 3(1), 1-15. https://www.researchgate.net/publication/323225661

Azizah, N., & Kuswanto, H. (2020). Item Analysis of Multiple-choice Questions (MCQs) Assessment Tool For Quality Assurance Measures. Journal of Physics: Conference Series, 1657(1), 012086. https://www.researchgate.net/publication/355098496

Budiarto, M. T., & Pramudya, I. (2017). Item Analysis of Multiple Choice Questions: Assessing an Assessment Tool in Medical Students. Journal of Educational Research and Evaluation, 21(2), 148-153. http://jurnal.unissula.ac.id

Center for Excellence in Learning and Teaching. (n.d.). Guide to Item Analysis Reports. Schreyer Institute for Teaching Excellence at Penn State University. https://www.schreyerinstitute.psu.edu/pdf/GuideToItemAnalysis.pdf

Djamari, M (2017). *Pengukuran, Penilaian dan Evaluasi Pendidikan.* Parama Publhising : Yogyakarta.

Hidayatullah, R., & Widiastuti, S. (2020). Pengembangan Kurikulum Pendidikan Agama Islam (*PAI*) di Madrasah. Nadwa: Jurnal Pendidikan Islam, 14(2), 309-328. https://journal.walisongo.ac.id/index.php/Nadwa/article/download/15384/4738

Nurhayati, N., & Asrori, M. (2020). The Development of Islamic Education Curriculum in Madrasah Based on the 2013 Curriculum in Indonesia: A Case Study in Madrasah Aliyah Negeri 1 Semarang Central Java Indonesia. Nadwa: Jurnal Pendidikan Islam, 14(2), 329-348. https://journal.walisongo.ac.id/index.php/Nadwa/article/download/15385/4739

Prasetyo, B., & Jumadi, J. (2020). Developing Critical Thinking Skills Assessment Instrument Based on Revised Bloom's Taxonomy Using the Rasch Model. Journal of Research and Advances in Mathematics Education, 5(2), 139-150. https://journal.uny.ac.id/index.php/reid/article/download/43672/17723