Jurnal Pendidikan Agama Islam
ISSN: 1829-5746 | EISSN: 2502-2075
Vol. 19, No. 2, Desember 2022
DOI: https://doi.org/10.14421/jpai.2022.192-09

Utilizing Item Response Theory Approach for Calibrating Items in the Final Assessment of Islamic Cultural History Subject

Rani Putri Prihatin[1]✉, Siti Fatonah[1], Iqbal Faza Ahmad[2]
[1]Universitas Islam Negeri Sunan Kalijaga Yogyakarta, Indonesia
[2]Universitas Negeri Yogyakarta, Indonesia

ABSTRACT

Purpose –The aim of this exploratory descriptive study with a IRT approach is to evaluate the quality of items used in the final assessment of 10th grade Islamic cultural history subject at Madrasah Aliyah Negeri (MAN) 2 Bantul Yogyakarta.

Design/methods/approach–Data was collected through documentation of student responses at the end of the year assessment. Student response data were analyzed using the item response theory approach.

Findings – Based on the analysis of 25 questions on the Islamic Cultural History subject at MAN 2 Bantul, the study concludes that the instrument fits the 1 PL model (Logistics parameter), as indicated by the results of the model fit test. The parameter analysis of item difficulty level with the 1 PL model reveals that 7 items fall under the very easy category, 5 items under the easy category, 9 items under the medium category, and 4 items under the very difficult category. Notably, none of the items fall under the difficult category in this 1 PL model.

Research implications/limitations –The findings on the distribution of item difficulty levels could guide educators in revising and improving the quality of assessment tools, ensuring that the questions are well-balanced and challenging enough to effectively evaluate students' knowledge and skills. Moreover, the absence of items falling under the difficult category in the 1 PL model suggests a need for more rigorous and challenging items to enhance the validity and reliability of the assessment instrument.

Originality/value – The research value of this study lies in its contribution to the development of a valid and reliable assessment tool in the field of Islamic cultural history subject.

OPEN ACCESS

ARTICLE HISTORY
Received: 2-12-2022
Revised: 28-12-2022
Accepted: 31-12-2022

KEYWORDS
History of Islam; Religious Education; Item Response Theory

Introduction

The final assessment is an activity carried out by the education unit at the end of the even semester to measure the achievement of student competence at the end of the even semester in the education unit using the package system (Kemendikbud, 2017). In an

CONTACT: ✉ rani.putri289@gmail.com

ideal assessment, the learning outcomes test instrument prepared by the teacher can provide a lot of information about students' abilities. However, many problems that arise when the teacher composes the test instrument, it turns out that there are still many weaknesses in the preparation process, resulting in a test that does not have a valid measuring power. A test instrument that does not have a valid measuring power will not be able to provide any information about the test taker's ability. To evaluate the achievement of the learning process that is commonly done by teachers is to develop a learning outcome test instrument (Prihatin & Hamami, 2022).

Basically, doing item analysis is very important for teachers to do. It is intended that teachers can find out the reliability and validity of the questions they have compiled (Kheyami et al., 2018). In addition, through item analysis, you will get feedback which can later be used as a reference in making decisions. Questions that fall into the proper category can be used immediately, questions in the less feasible category can be revised, while questions in the unfit category can be dropped so they must be deleted or replaced with other questions (Karkal & Kundapur, 2016). But unfortunately, the knowledge and skills of teachers in conducting item analysis are still relatively low (Sumiati et al., 2018). Most teachers tend to make questions according to the core competencies and basic competencies that have been determined without first measuring whether students have understood the materials to be tested. Thus, it is certain that student competence cannot be measured accurately (Kurniawan et al., 2017).

Schools and teachers become an important part in providing facilities in the student learning process, without a teacher, students will not develop and improve their learning outcomes. One of the efforts that teachers make to determine the increase in student knowledge is by conducting an assessment or also known as an assessment system for student learning outcomes while in Madrasah. Learning outcomes are carried out to measure understanding and mastery of the material provided by the teacher to students, which means that this assessment can provide an overview of each student in achieving their competence (Perdana, 2018). Istiyono explained that Assessment is the process of gathering information about students and classes for instructional decision-making purposes (Istiyono, 2018). To conduct an assessment, an instrument is needed that can measure students' abilities accurately. In the case of summative or end-of-semester exams, it is usually done with multiple choice. The process of assigning a number to something or someone based on a rule is called measurement. The activity of systematically determining numbers for an object is defined as a measurement by Mardapi (2012). Therefore the assessment is very important to be carried out by the teacher, by conducting an assessment the teacher can find out the abilities of the students, in carrying out this activity one way is to use test and non-test questions. These test questions are often used in schools or madrasas to measure students' abilities in carrying out assessments. The test questions are considered able to accommodate every material that has been conveyed by the teacher to students in each lesson in both description and multiple choice questions (Perdana, 2018).

Yahya Qohar as quoted by Badrun Badrun said that an evaluation must meet 6 (six) requirements, namely: (1) Reliable; (2) Valid; (3) Objective; (4) Must be Discriminatory; (4) Imperfective; and (5) Easy to Use. Discriminatory means that he is able to distinguish the value and size of students in their achievement of the subject matter that has been given. And in line with this opinion, namely that an evaluation must meet discriminatory requirements, a test must be analyzed in terms of the level of difficulty and distinguishing power (Kartowagiran, 2009).

Syamsudin said that there are 4 (four) ways that can be used to assess a good test, namely: (1) examine honestly the questions that have been prepared; (2) conduct a question analysis (item analysis); (3) checking validity (checking validity); and (4) checking reliability. The first method is a subjective method, while the other 3 (three) can be carried out quantitatively and measurably. Problem analysis can be done by: (1) Difficulty level (difficulty level); (2) Distinguishing Power (Discriminative test); and (3) the pattern of answers to questions (Wening, 2012).

According to Azis, the level of difficulty is how easy or difficult an item is for a group of students. In general, it can be said that the level of difficulty is the level of easy or not a question given to a group of students. Purwanto explained that the distinguishing power (DB) is the ability of the THB (Learning Outcome Test) items to distinguish students who have high and low abilities (Purwanto, 2010). This discriminatory analysis aims to determine the ability of the questions in distinguishing students who are classified as capable (high in achievement) and students who are classified as weak in achievement (Sudjana, 2009).

While the pattern of answers to questions is the distribution of the testee in terms of determining the answer choices on multiple choice questions. The pattern of answers to questions is obtained by counting the number of testees who choose the answer choices a, b, c, or d or who do not choose any option (blank). In evaluation terms it is called omit, abbreviated as O. From the pattern of answer questions, it can be determined whether the distractor functions as a distractor properly or not. A distractor not selected by the testee at all means that the distractor is ugly, too glaringly misleading. On the other hand, a distractor (distractor) can be said to function properly if the distractor has great appeal for test takers who do not understand the concept or master the material.

Assessment of learning outcomes by educators is carried out continuously to monitor the process, progress and improvement of results in the form of daily tests, midterm exams, end-of-semester exams and class promotion exams. Assessment of learning outcomes by students is used to assess the achievement of graduate competency standards for all subjects. Assessment of learning outcomes by the government in the form of a national exam aims to assess the achievement of graduate competencies for all subjects (Sudaryono, 2011). Assessment aims to determine the characteristics of an object to be measured. In particular, the measurement of education includes the measurement of learning outcomes that cover various fields, depending on the object of learning outcomes that you want to measure. It can be said that the measurement is quantitative,

to determine the value quantitatively a measuring instrument is needed, one of which is a test. The measuring instrument or test instrument commonly used in evaluating student learning outcomes is a set of questions.

In general, assessment tools (instruments) can be categorized into two forms, namely: 1) Tests; and 2) Not Test (non-test). The measurement tools included in the non-test category are: a) Questionnaire; b) Interview; c) Match List (check list); d) Observation or Observation; e) Assignment; f) Portfolio; g) Journal; h) Inventory; i) Self Assessment (Self Assessment); j) Assessment by friends (peer assessment). While the test is a number of questions that must be answered, or questions that must be selected or responded to, tasks that must be done by the person being tested at a certain time. A test is a number of questions that have true or false answers, questions that require answers or are given a response to measure a person's level of ability in certain aspect (Wening, 2012).

In relation to the preparation of a test, Badrun Kartowagiran explained the steps for preparing a test in general, namely: (1) Determining the purpose of the test; (2) the arrangement of the grid; (3) Writing questions; (4) Questions review (Review and Revision of Questions); (5) test questions; (6) assembling questions into test kits. Determination of goals can be in the form of specific goals, namely seeing the level of achievement of a program. The question grid is a description of the scope of the questions to be tested and provides details about the questions needed in the test. Writing questions is a description of the type and level of behavior to be measured into questions whose characteristics are in accordance with the details in the grid. A question study is a theoretical study of questions. Testing questions is a step to determine the quality of the questions to be tested by looking at the responses of test takers empirically. Assembling questions is presenting questions into an integrated test kit (Kartowagiran, 2009).

Measuring tools that can be used in the process of evaluating the learning process can be in the form of homework assignments, quizzes, midterm exams (UTS), and final semester exams (UAS). The test is a form of instrument used to measure a number of questions that have true or false answers, or all true or partially true with the aim of knowing the learning achievements or competencies that have been achieved by students in certain fields (Djemari, 2008) Summative Assessment or also known as PAS (Semester Final Assessment) is sometimes made too difficult or too easy so that it is difficult for educators to distinguish students' abilities. Therefore, it is necessary to test/analyze the test questions in the hope that the results obtained accurately reflect the students' abilities (Istiyono et al., 2020).

The questions developed in UAS are generally in the form of object tests (multiple choice) in multiple choice questions divided into two systems, namely statements (*stem*) and alternative answers (*option*). Stem questions are usually in the form of statements or questions, while option questions are several of the answer choices. One of the alternative answer choices is the answer key, while the others are referred to as distractors (*distractors*). A good question must have a relatively homogeneous distractor, so it is not

easy for students to guess (Ratnaningsih & Isfarudi, 2013). This theory is where the classical test theory or referred to as *Classical test Theoty* (CTT) in the process of conducting item analysis, the advantages of this theory are that the level of difficulty and discriminatory power of items can be measured manually (Nurcahyo, 2016). If the test is difficult, it means that the level of student ability is low, in other words, if the test is tested on participants with low ability, then the level of difficulty of the test will be high. On the other hand, if the test is easy, it means that the level of students' ability is high, in other words, the level of difficulty of the questions will be low when the test is tested on the high-ability group. The second level of difficulty of the questions is defined as the proportion of students in the group who answered the questions correctly (Hambleton et al., 1985).

Hambleton, Swaminathan, & Rogers stated that classical test theory has indeed dominated and is widely used in the world of measurement in recent years, almost all the concepts of validity and reliability that are known today are developments of classical test theory, but in this case classical test theory has several limitations.. Therefore began to develop the theory of item response or *Item Response Theory* (IRT) which is a theory that was developed to improve the limitations that exist in classical test theory. Item response theory is one way to assess item feasibility by comparing the average item performance against the performance evidence predicted by the model, the main purpose of the item response theory being developed is to overcome the weakness of classical test theory which is not independent of the sample (Hambleton et al., 1985).

There are 3 models of item response theory, the measurement model is based on the number of item parameters entered into the model, namely the one-parameter model (1PL) is (b) the level of difficulty of the items, the two-parameter model (2PL) is (b) the difficulty level of the items. and (a) distinguishing power, the three-parameter model is (b) the level of difficulty of the items and (a) distinguishing power (g) pseudo guess (Retnawati, 2014).

Arikunto (2008) stated that item analysis has several benefits including: 1) identifying good and bad items, 2) being able to obtain information to improve the test used both in terms of content and constructs, 3) getting an overview of the state of the tests that are prepared.

The basic concept of item response theory is that the subject's performance on a test can be predicted or explained by a set of factors called traits, latent traits or abilities and the relationship between the subject's performance on an item and a set of underlying latent abilities can be described by a monotonically increasing function. which is referred to as the item characteristic curve (*Item characteristic curve-ICC*) (Retnawati, 2014).

Understanding *Item Response Theory* (IRT) assesses a person's behavior can be explained by the characteristics of the person concerned to certain limits (Djemari, 2008). Hambleton and Swaminathan stated that item response theory is one way to assess item eligibility by comparing the average item appearance to the appearance of evidence of

group ability (Ratnaningsih & Isfarudi, 2013). *Item Response Theoty* (IRT) has the advantage that *not group dependent and not item dependent.* So that the natural difficulties when using classical res theory can be overcome by the IRT statistical method which estimates students with different ability levels (Hambleton et al., 1985).

Item response theory emphasizes the probability of a test taker's correct answer, item parameters and test taker's parameters being linked through a mathematical function or a mathematical formula model. In this formula, the probability of test takers answering the questions is understood as a logistic function of the difference in parameters entered into the model.

The subject of Islamic Cultural History at Madrasah Aliyah is one of the subjects that examines the origin, development, role of Islamic culture/civilization in the past, starting from the da'wah of the Prophet Muhammad in the Mecca and Medina periods, the leadership of the people after the Prophet SAW died, until the development of Islam in the classical period (the golden age) in 650 AD – 1250 AD, the medieval/regressive era (1250 AD–1800 AD), and the modern period/revival era (1800-present), as well as the development of Islam in Indonesia and in the world.. Based on the explanation above, learning Islamic Cultural History (SKI) is essentially a knowledge transfer activity carried out by teachers to students that is closely related to past events, be it political, social, or economic events that happened in a certain situation. Islamic state and experienced by the Islamic community.

The preparation of the items for the final assessment of SKI Subjects for class X MAN 2 Bantul is carried out by the tutor himself. This item has never been analyzed by the school, namely the SKI subject teacher, so the quality of the item is not known. Meanwhile, according to the description of the previous explanation, it can be seen that the learning evaluation, namely the test, plays a very important role in the follow-up decision to students regarding the achievement of learning objectives. The problem that will be studied by the researcher is departing from the assessment aspect using questions that are designed so well that it raises the question, whether the assessment is in accordance with the ability of the students who answer, this is also related to the level of validity of the test questions, namely the extent to which the test correctly measures aspects to measure (Sudaryono, 2011).

Assessment is a process of interpreting all information in a systematic, periodic, comprehensive manner related to the process and results of growth and development achieved by students through teaching and learning activities, and interpreting this as an assessment decision (Oktarina & Fatonah, 2021). In this case, the researcher used a sample at MAN 2 Bantul, Yogyakarta. This study aims to determine the quality of the year-end test questions for Class X SKI at MAN 2 Bantul, Yogyakarta.

Methods

This research is an exploratory descriptive study with a quantitative approach (Yusron et al., 2020). Data collection was carried out through documenting student responses in the 2020/2021 The final assessment (PAT) for the subject of Islamic cultural history for grade 10 Madrasah Aliyah. The final assessment Questions for the SKI subject at MAN 2 Bantul consist of 25 questions in the form of multiple choice. While the number of respondents is 157 students.

The research steps include: (1) Collecting answer student from school databases; (2) preparing data to be analyzed in Microsoft Excel format; (3) To do analysis data with R, and serve the result; (5) Interpreting the data from the analysis. Analysis conducted in several Step. On Step first empirical validity analysis was performed. In the second stage, an analysis of the characteristics of the test was carried out based on the item response theory and the fit of the logistic parameter model was analyzed. This analysis process uses the help of the R Program (Syafii et al., 2021).

Result and Discussion

1. Empirical Validity Test

Product Moment Validity Test with SPSS The data used for the validity test is the answer to the final exam for the 2020/2021 school year, with a total of 157 students or N = 157 and 25 items. The calculated r value of the SPSS calculation results was then consulted with the r table, then the decision was made by comparing the calculated r value with the r table, with an r table value of 0.1557. Using the following statement (Setemen, 2018):

- If the value of r count > r table, then the item in the questionnaire is declared valid.
- If the value of r count < r table, then the item in the questionnaire is declared invalid.

Table 1. Empirical Validity

Item	Rcount	Rtable	Information
1	0.538	0.1557	Valid
2	0.519	0.1557	Valid
3	0.553	0.1557	Valid
4	0.026	0.1557	Invalid
5	0.562	0.1557	Valid
6	0.466	0.1557	Valid
7	0.424	0.1557	Valid
8	0.515	0.1557	Valid
9	0.496	0.1557	Valid
10	0.547	0.1557	Valid
11	0.482	0.1557	Valid
12	0.593	0.1557	Valid
13	0.277	0.1557	Valid
14	0.453	0.1557	Valid
15	0.434	0.1557	Valid
16	0.377	0.1557	Valid
17	0.425	0.1557	Valid
18	0.261	0.1557	Valid

19	0.454	0.1557	Valid
20	0.620	0.1557	Valid
21	0.511	0.1557	Valid
22	0.525	0.1557	Valid
23	0.518	0.1557	Valid
24	0.387	0.1557	Valid
25	0.469	0.1557	Valid

From the results of the empirical validity table above, it can be concluded that the SKI questions at the end of class X for the 2020/2021 academic year stated 24 valid questions and 1 invalid question.

2. Model Fit Test

The goodness of fit of the model can be determined by comparing the chi-square calculation results and the chi-square table with certain degrees of freedom. An item is said to fit a model if the calculated chi-square value does not exceed the chi-square value of the table. The goodness of fit of the model can also be seen from the probability or significance value, if the sig value is less than the alpha value, then the item is said to not fit the model.

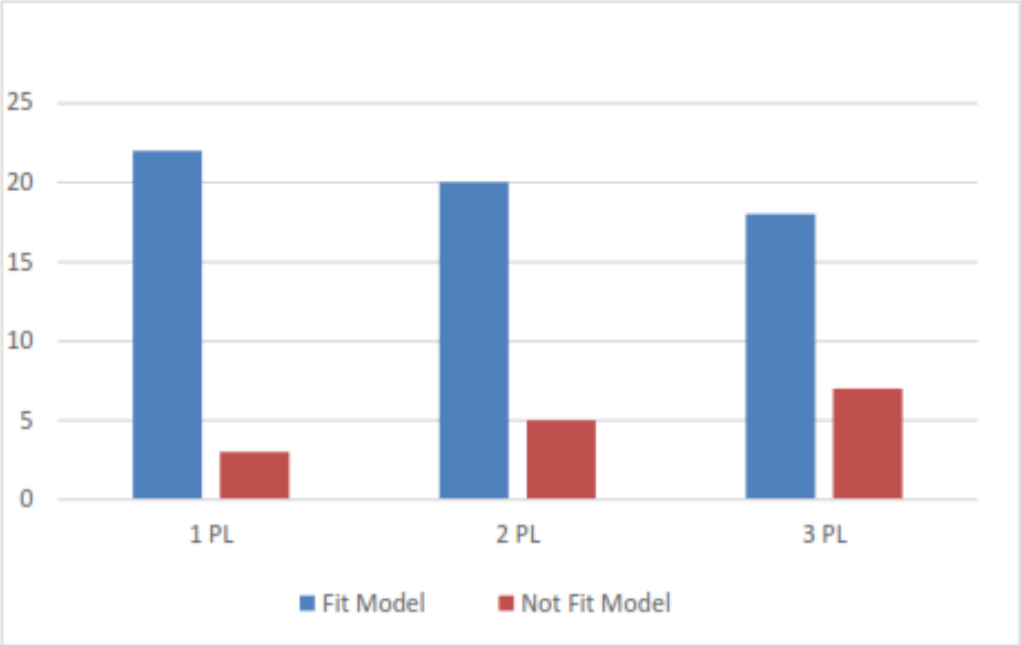
Table 2. Model Fit Test

Items	1PL		2PL		3PL	
	pvalue	Category	pvalue	Category	pvalue	Category
Item_1	0.17	Fit	0.05	Fit	0.02	Not fit
Item_2	0.15	Fit	0.30	Fit		Not fit
Item_3	0.45	Fit	0.54	Fit	0.73	Fit
Item_4	0.00	Fit	0.10	Fit	0.03	Not fit
Item_5	0.36	Fit	0.25	Fit	0.18	Fit
Item_6	0.36	Fit	0.33	Fit	0.01	Not fit
Item_7	0.10	Fit	0.07	Fit	0.14	Fit
Item_8	0.51	Fit	0.32	Fit	0.26	Fit
Item_9	0.48	Fit	0.89	Fit	0.58	Fit
Item_10	0.09	Fit	0.41	Fit	0.20	Fit
Item_11	0.28	Fit	0.15	Fit	0.14	Fit
Item_12	0.36	Fit	0.58	Fit	0.46	Fit
Item_13	0.05	Fit	0.19	Not fit	0.55	Fit
Item_14	0.94	Fit	0.88	Fit	0.87	Fit
Item_15	0.37	Fit	0.52	Fit	0.39	Fit
Item_16	0.01	Not fit	0.17	Fit	0.11	Fit
Item_17	0.04	Not fit	0.04	Not fit	0.05	Fit
Item_18	3.39	Fit	0.03	Not fit	0.14	Fit
Item_19	0.22	Fit	0.29	Fit	0.53	Fit
Item_20	0.11	Fit	0.61	Fit	0.48	Fit
Item_21	0.23	Fit	0.35	Fit	0.58	Fit
Item_22	0.02	Not fit		Not fit		Not fit
Item_23	0.07	Fit		Not fit		Not fit
Item_24	0.40	Fit	0.28	Fit	0.12	Fit
Item_25	0.55	Fit	0.58	Fit	0.01	Not fit
Total Matches		22		20		18

In the model goodness of fit test that was carried out, the model Fit for use in the analysis of the year-end items for the SKI class X year 2020/2021 used 1 parameter. This is because in each item more matches are found using the 1-parameter model because the results are greater than the alpha value, which is 0.05. Of the 25 questions, only three questions were found that were Not fit because the significant value was below the alpha value, which was 0.05. As in the table below, the results of the fit of the 1 parameter, 2 parameter and 3 parameter models are found.

Based on the results in the 3 tables above, it turns out that the model that produces the most items that match the model is the 1-parameter logistic model, this means that the 1-parameter model is the model that can be chosen for item analysis. Paying attention to the results of the end-of-year exam data for class X SKI in the 2020/2021 school year, it was found that 22 of the 25 items fit the 1-parameter model.

This is due to the response data used for analysis as many as 157 test participants, the more responses the test takers used, the greater the chi-square value of the count. The greater the acquisition of the chi-square value, the greater the chance of rejecting the hypothesis that the item is Fit for analysis with the 1-parameter model (Retnawati, 2014).



Figures 1. Model Fit

3. Parameter Estimation of 1PL Model Items

The 1 PL parameter model is related to the level of difficulty of the test items that the level of item difficulty is often associated with the ability of the respondents with a moderate level of item difficulty and there are easy items, the item difficulty level is a series from easy to difficult. Items that are too easy give good results so that they are classified as good categories and items that are too difficult will also give bad results and of course

there are categories that are not good for respondents (Hasnah, 2017). The value of b or the difficulty of a good item ranges from -2 to 2, a value closer to the negative line indicates that the item is too easy, as well as the closer to the positive line indicates the item is getting more difficult (Hasnah, 2017).

Table 3. Estimation of 1PL Model Item Parameters

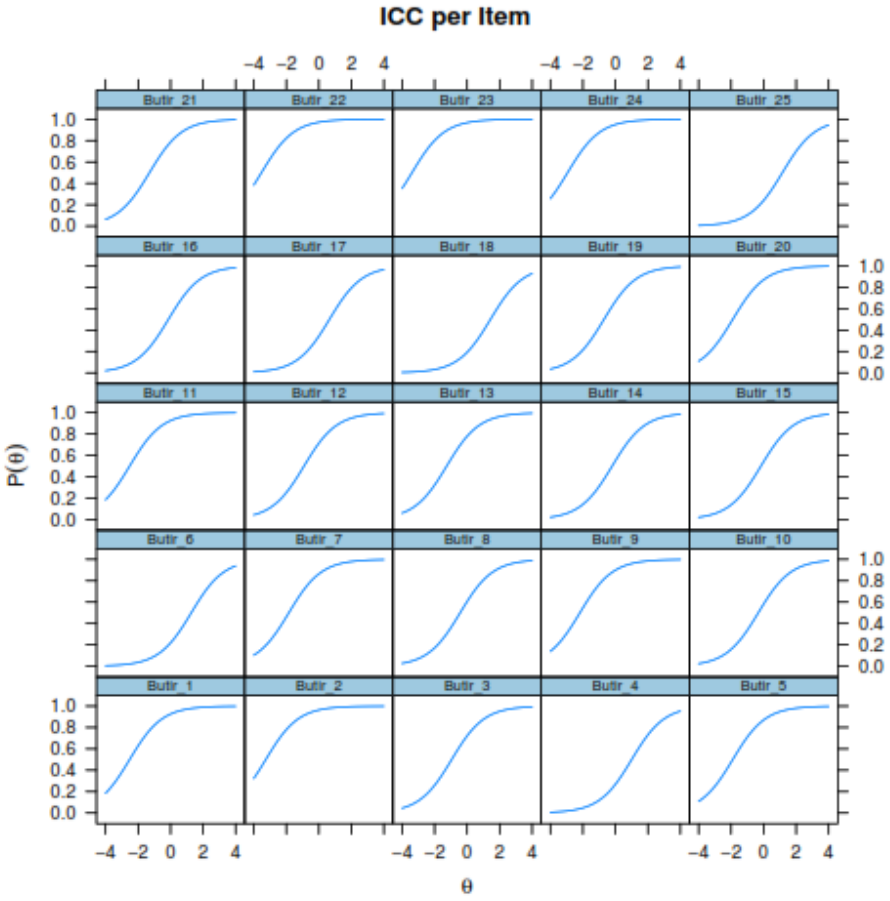
Items	a	b	G	U	Category
Item_1	1	-2.51923	0	1	Very easy
Item_2	1	-3.2643	0	1	Very easy
Item_3	1	-0.95017	0	1	Moderate
Item_4	1	1.003429	0	1	Hard
Item_5	1	-1.8996	0	1	Easy
Item_6	1	1.271006	0	1	Hard
Item_7	1	-1.84736	0	1	Easy
Item_8	1	-0.42882	0	1	Moderate
Item_9	1	-2.1835	0	1	Very easy
Item_10	1	-0.32975	0	1	Moderate
Item_11	1	-2.51923	0	1	Very easy
Item_12	1	-0.91351	0	1	Moderate
Item_13	1	-1.25844	0	1	Easy
Item_14	1	-0.19915	0	1	Moderate
Item_15	1	-0.16669	0	1	Moderate
Item_16	1	-0.13429	0	1	Moderate
Item_17	1	0.650765	0	1	Moderate
Item_18	1	1.435148	0	1	Hard
Item_19	1	-0.70011	0	1	Moderate
Item_20	1	-1.8996	0	1	Easy
Item_21	1	-1.29922	0	1	Easy
Item_22	1	-3.53054	0	1	Very easy
Item_23	1	-3.39095	0	1	Very easy
Item_24	1	-2.94059	0	1	Very easy
Item_25	1	1.15365	0	1	Hard

The results of the analysis can be explained that *b* explain the level of difficulty of the test items and *a* explain the differentiating power of the question. Obtained value *a* (difference power) the the same because the analysis used is an analysis of 1 Logistics Parameters. The estimation results are then categorized based on the range as follows:

Table 4. Item Parameter Category 1PL

Parameter	Classification	Category	Number of Items
Level Difficulty (b)	b < -2.00	Very easy	7
	-2.00 b < -1.00	Easy	5
	-1.00 b < 1.00	Moderate	9
	1.00 b 2.00	Hard	4
	b > 2.00	Very Difficult	0

Based on the category table above, it can be seen that the distribution of the level of difficulty of the 25 items is quite balanced. This can be understood from the proportion of very easy items totaling 7 items and easy items totaling 5 items. while the items with the level of difficulty are 4 items. For items in the medium category as many as 9 items. However, the distribution of difficulty levels should be increased. This can be done by increasing the types of questions that have difficult or very difficult categories.



Figures 1. ICC per Item

Based on the 1PL model ICC graph, it can be seen that the distribution of difficulty levels is quite good. The ability that must be possessed by students to answer 25 test questions for the Islamic Cultural History at MAN 2 Bantul with a probability of 50% is in the minimum range of abilities -2.5 to 2.8.

The findings of this study suggest that the 1 PL model is suitable for assessing the Islamic Cultural History subject at MAN 2 Bantul. The analysis of item difficulty level indicates that the majority of the questions fall under the easy and medium categories, while none of the items fall under the difficult category. This implies that the test may not be challenging enough for students who are well-versed in the subject matter. However, it is important to note that this study only analyzed one particular test administered at MAN 2 Bantul and the findings may not be generalizable to other schools or contexts.

Further research could explore the reasons why none of the items fall under the difficult category and whether this is indicative of a larger issue with the curriculum or teaching methods. Additionally, future studies could investigate how to increase the difficulty level of the assessment without sacrificing its validity and reliability. This study could serve as a starting point for researchers and educators who are interested in improving the quality of assessments in Islamic Cultural History or other related subjects.

Conclusion

In conclusion, this study provides evidence that the instrument used to assess students' knowledge and skills in the Islamic Cultural History subject at MAN 2 Bantul fits the 1 PL model, indicating its validity and reliability. The parameter analysis of item difficulty level further reveals the distribution of item difficulty levels, which can inform curriculum developers and educators in designing and refining assessment tools that accurately measure students' learning outcomes. The finding that none of the items fall under the difficult category in this 1 PL model may suggest that further investigation is needed to ensure that the assessment tool is sufficiently challenging for students of varying abilities. Overall, this study highlights the importance of using appropriate item response theory models in calibrating assessment instruments and provides practical insights for improving the quality of education in the field of Islamic cultural history.

References

Arikunto, S. (2008). *Penelitian tindakan kelas/Suharsimi Arikunto*. Bumu Aksara.

Djemari, M. (2008). *Teknik penyusunan instrumen tes dan nontes*. Mitra Cendekia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985). *Principles and applications of item response theory*. Kluwer-Nijhoff Publishing Company.

Hasnah, H. (2017). Analisis kualitas soal matematika Ujian Sekolah kelas XII IPA SMA Negeri di Watansoppeng berdasarkan Teori Respon Butir. *PEP Educational Assessment*, *1*(1), 27–34.

Istiyono, E. (2018). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika dengan Teori Tes Klasik dan Modern*. UNY Press.

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of Computerized Adaptive Testing to Measure Physics Higher Order Thinking Skills of Senior High School Students and Its Feasibility of Use. *European Journal of Educational Research*, *9*(1), 91–101.

Karkal, Y. R., & Kundapur, G. S. (2016). Item analysis of multiple choice questions of undergraduate pharmacology examinations in an International Medical School in India. *Journal of Dr. NTR University of Health Sciences*, *5*(3), 183.

Kartowagiran, B. (2009). *Pengantar Teori Tes Klasik*. Pascasarjana UNY & Dinas Pendidikan

Propinsi Yogyakarta.

Kemendikbud. (2017). *Panduan Penilaian oleh Pendidik dan Satuan Pendidikan untuk Sekolah Menengah Pertama*. Kementerian Pendidikan dan Kebudayaan Direktorat Jenderal Pendidikan Dasar dan Menengah Direktorat Pembinaan Sekolah Menengah Pertama.

Kheyami, D., Jaradat, A., Al-Shibani, T., & Ali, F. A. (2018). Item analysis of multiple choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal*, *18*(1), e68.

Kurniawan, R. Y., Prakoso, A. F., Hakim, L., Dewi, R. M., & Widayanti, I. (2017). Pemberian Pelatihan Analisis Butir Soal Bagi Guru di Kabupaten Jombang: Efektif? *Jurnal Pemberdayaan Masyarakat Madani (JPMM)*, *1*(2), 179–193.

Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Nuha Medika.

Nurcahyo, F. A. (2016). Aplikasi IRT dalam analisis aitem tes kognitif. *Buletin Psikologi*, *24*(2), 64–75.

Oktarina, A., & Fatonah, S. (2021). Pengamatan Tentang Pembelajaran Dan Penilaian Pada Anak Usia Dini Di Era Pandemi Covid-19. *Cakrawala Dini: Jurnal Pendidikan Anak Usia Dini*, *12*(1), 31–40.

Perdana, S. A. (2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan rasch model. *Jurnal Kiprah*, *6*(1), 41–48.

Prihatin, R. P., & Hamami, T. (2022). Learning Assessment Model for Islamic Religious Education. *Nusantara: Jurnal Pendidikan Indonesia*, *2*(2), 373–390.

Purwanto. (2010). *Evaluasi Hasil Belajar*. Pustaka Pelajar.

Ratnaningsih, D. J., & Isfarudi, I. (2013). Analisis butir tes objektif ujian akhir semester mahasiswa Universitas Terbuka berdasarkan teori tes modern. *Jurnal Pendidikan Terbuka Dan Jarak Jauh*, *14*(2), 98–109.

Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.

Setemen, K. (2018). Pengembangan Dan Pengujian Validitas Butir Instrumen Kecerdasan Logis-Matematis. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *15*(2).

Sudaryono, S. (2011). Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan Dan Kebudayaan*, *17*(6), 719–732.

Sudjana, N. (2009). *Penilaian Hasil Proses Belajar Mengajar*. Remaja Rosda Karya.

Sumiati, A., Widiastuti, U., & Suhud, U. (2018). Workshop Teknik Menganalisis Butir Soal dalam Meningkatkan Kompetensi Guru di SMK Cileungsi Bogor. *Jurnal Pemberdayaan*

Masyarakat Madani (JPMM), *2*(1), 136–153.

Syafii, A., Haryanto, H., Ahmad, I. F., & Fauziah, A. (2021). Analysis of Items with Item Response Theory (IRT) Approach on Final Assessment for Al-Quran Hadith Subjects. *Jurnal Pendidikan Agama Islam*, *18*(1), 167–194.

Wening, S. (2012). *Materi Evaluasi Pembelajaran*. Fakultas Teknik Universitas Negeri Yogyakarta.

Yusron, E., Retnawati, H., & Rafi, I. (2020). Bagaimana hasil penyetaraan paket tes USBN pada mata pelajaran matematika dengan teori respon butir? *Jurnal Riset Pendidikan Matematika*, *7*(1), 1–12.