Psychology, Evaluation, and Technology in Educational Research
6 (2), 2024, 107-118

Available Online: http://petier.org/index.php/PETIER

Analysis of numeracy skills in Islamic Boarding Schools: Gender bias

Rosid Bahar [1 a] *, Ahmad Firdaus [2 b]

[1] Sekolah Tinggi Agama Islam Al-Andina. Jl. Raya Selakopi, Sukabumi 43155 Indonesia

[2] Sekolah Tinggi Agama Islam Al-Masthuriyah. Jl. Raya Sukaraja-Sukabumi, Sukabumi, 43155 Indonesia

[a] rosidbahar@gmail.com; [b] firdaus.ahmad1st@gmail.com

* Corresponding Author.

Received: 19 December 2023; Revised: 4 January 2024; Accepted: 12 January 2024

Abstract: Numeracy skills are an important point in the structure of mathematics, and boarding school students are no exception. This study aims to identify gender bias in numeracy assessment in Islamic boarding schools. This research is a descriptive exploratory research using quantitative methods. The instrument used in this study was a numeracy test of 25 questions consisting of matching, multiple-choice, complex multiple-choice, and description questions. The research subjects involved 383 students in West Java consisting of 4 pesantren in 4 cities, namely West Bandung Regency, Cirebon Regency, Tasikmalaya Regency, and Tasikmalaya City. Quantitative analysis used Item Response Theory (IRT) followed by Differential Item Functioning (DIF) analysis with the Mantel-Haenzel method. The results showed that the instrument was suitable for use because it met the standards of validity and reliability. The model fit test that meets is GPCM, and DIF analysis shows that there is 1 number, namely number 21, in numeracy questions that indicate gender bias. The results of this analysis indicate that the numeracy test instrument set is suitable for use by boarding school students with minimal gender bias.

Keywords: Islamic Boarding School, Numeracy Skill, Differential Item Functioning

How to Cite: Bahar, R., & Firdaus, A. (2024). Analysis of numeracy skills in Islamic Boarding Schools: Gender bias. *Psychology, Evaluation, and Technology in Educational Research*, *6*(2), 107-118. https://doi.org/10.33292/petier.v6i2.194

INTRODUCTION

Islamic Boarding Schools are the oldest educational institutions and are among the institutions that have played a role in fighting for the independence of the Republic of Indonesia (Masqon, 2014; Muafiah et al., 2022). The history also records that the establishment of Islamic boarding schools was a place of study for Muslim students (Santri) to focus on Islamic religious learning so that they have a willing to become clerics of the religion in their area (Isbah, 2020; Wekke & Hamid, 2013).

The current development of Islamic boarding schools has motivated parents from various circles to send their children to these institutions. This motivation could arise from the parents, either alumni of an Islamic boarding school or parents who want to choose this school that also provides school education (Supriatna, 2018). These motivations have an impact on student input in increasingly diverse backgrounds in Islamic boarding school institutions. One of the most prominent problems regarding the impact of student input is that many students still lack attention to general subjects, such as mathematics (Ramdhani et al., 2021; Yusnita, 2011).

The current development of Islamic boarding schools has also encouraged the birth of Law No. 18 of 2019 concerning Islamic boarding schools, where Islamic boarding schools are not

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

10.33292/petier.v6i2.194

only about teaching Islamic religion but have a more role in upholding the true teachings of Islam, which are reflected in the character of tolerance, balance and community empowerment (Ministry of Religion, 2020).

Islamic boarding schools are formal institutions that are on par with other schools, starting from elementary to higher education. According to Minister of Religion Regulation (PMA) Number 31 of 2020, the typology of Islamic boarding schools has also changed, including Formal Diniyah Education (PDF), Muadalaah (Muallimin & Salafiyah), and the highest at Ma'had Aly. The levels were changed to Ula (Primary), Wustha (intermediate), Ulya (Upper), and Ma'had Aly (Higher Education). In technical learning, the provisions in the law and PMA also apply a minimum of 5 general subjects, including mathematics, Pancasila and citizenship education (PPKN), Indonesian language, natural sciences, and social sciences (Ministry of Religion, 2020).

This regulation seems to be a concern to begin to know the continuity of the implementation of this Law, including in terms of proficiency in general subjects such as mathematics. Moreover, it is a fact that classroom conditioning in Islamic boarding schools uses rules of separation between male and female students, which leads to gender bias. This kind of culture is commonplace in Islamic boarding school environments. This culture is strongly attached to them today and will probably continue to be preserved. This process is understandable because Islamic boarding schools still maintain and implement fiqh practices following Islamic teachings (Sahri & Hidayah, 2020). The question is, is it possible that gender bias will occur in the process of teaching and learning activities, especially in assessments?

The initial technique for implementing the Law and Minister of Religion Regulations (PMA) can start by constructing instruments and analyzing them to obtain an instrument construct that can minimize gender bias. In the study, the mathematical instruments in Islamic boarding schools are more specific to numeracy instruments.

Numeracy skills are an important point in the structure of mathematics, especially for children. It helps develop logical thinking skills, problem-solving skills, and skills and knowledge important to understanding the world around us (Whiteford, 2020). Numeracy skills are part of literacy skills, including the ability to identify, create, communicate, numerate, and use printed, written, and visual materials (Montoya, 2018).

Numeracy skills are included in the PISA indicators. Currently, it become a hot topic as Indonesia is ranked 72nd out of 78 countries (OECD, 2019). Research in Indonesia shows that 73% of secondary school students still lack an understanding of mathematical literacy (Ate & Lede, 2022). Moreover, there is no more specific research at the Islamic boarding school or Tsanawiyah level. This encourages research related to numeracy at the Islamic boarding school level. This point could be the beginning of discovering numeracy skills in Islamic boarding schools as a basis for implementing the Islamic Boarding School Law.

As the number of Islamic boarding schools and Santri is increasing yearly and a form of formal educational institution, it should always be active and dynamic in self-improvement through evaluation (Khaerudin & Munadi, 2020; Yasid, 2018). This action is carried out to maintain the implementation of learning programs in Islamic boarding schools in synergy with the outcomes that must be achieved by Indonesian education in meeting the demands of globalization. Evaluation is used to determine the feasibility of a program (Munthe, 2015). Evaluation aims to obtain recommendations on whether the learning program is good.

A component of evaluation that still does not have much attention by Islamic boarding schools is an assessment, which can measure students' success (Yasid, 2018). In other words, many Islamic boarding schools do not measure the success of their learning programs. In practice, the assessments attract a lot of attention, such as poor quality of the assessment

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

instruments in terms of content and constructs, which often lead to gender bias. This will certainly reduce the actual assessment, which should provide much information regarding implementing Teaching and Learning Activities (KBM) in Islamic boarding schools. Apart from assessment, gender bias in the world of education is often found in textbooks, language, and teacher-student interactions (Nadal, 2017). The impact of gender bias is quite detrimental to KBM, so it needs to anticipate this problem.

The impact of gender bias varies greatly, even leading to negative effects on students. Then, the possibility of gender bias in all elements of learning must be avoided, including learning in Islamic boarding schools that Islamic boarding schools have a large number of students. The assumption of gender bias must also be proven in the mathematics learning process in Islamic boarding schools. The previous research shows that Islamic studies, the method of memorizing the Al-Qur'an, have an influence on numeracy skills at the junior high school level. However, it has not measured them based on gender. Other research shows that single-sex learning has the highest value compared to mixed-sex learning processes (Franklin & Rangel, 2024). More specifically in mathematics, results from the Trends in International Mathematics and Science Study (TIMSS) (OECD, 2019) show high international differences in the gender gap in mathematics achievement, but not specifically in numeracy. Interestingly, countries with a high proportion of single-sex schools, such as Saudi Arabia, show unexpectedly high gains for girls (Basharat, 2022). On the other hand, other research shows no difference between single-sex learning and mixed classes (Clavel & Flannery, 2023). The research can be used as a reference for conducting further research regarding gender bias in mathematics in Islamic boarding schools. In the end, the research is expected to provide knowledge about the factors of gender bias and solutions to minimize gender bias.

METHODS

The research was exploratory, descriptive research with a quantitative approach. The subjects were 386 students in class 3 at the Ulya level of the Muadallah Muallimin Islamic Boarding School – An Islamic boarding school level equivalent to junior high school. The research was conducted in 4 Islamic boarding schools, spreading in three cities: Tasikmalaya City and Regency, West Bandung Regency, and Cirebon Regency, West Java, Indonesia. The number of subjects is the number of students from 21 classes with different class characteristics. The rationalization of subject selection is based on the Cluster Sampling Technique, namely by determining the sample if the object or data source is very broad. This sampling requirement is based on the population area which has been determined through two stages. The first stage determines the sample area, and the second determines the individuals in that area, which refers to cluster sampling (Berndt, 2020; Etikan, 2017; Rahman et al., 2022).

The instrument was a numeracy skill that relied on three main indicators: content, context, and process. The content relates to numbers, algebra, geometry and measurement, data, and uncertainty. Meanwhile, context is related to personal, socio-cultural, and scientific. Then, these three indicators were developed into a guideline containing numeracy content, competency achievement indicators, cognitive level, numeracy context, question indicators, question form, and question number.

Data was obtained from students' responses through a numeracy skills test in mathematics subjects. The skills test instrument is a numeracy test instrument consisting of 25 questions with three matching questions, eleven multiple choice questions, five complex multiple choice questions, and six essays. Each form of question has a different scoring, including matching

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

and multiple choice, which get a score of 1; complex multiple choice has a graded score of 1 to 4; and essays have a graded score of 1 to 5.

Data analysis used item response theory (IRT) Differential Item Functioning (DIF) model (Gomez-Benito et al., 2018; Hambleton et al., 1991; P. Lee & Joo, 2021; S. Lee & Kim, 2017; Raju, 1988; Retnawati, 2014). DIF analysis was carried out to see the gender bias in the questions using the item bias detection method to identify items with different functions for different groups. In other words, bias will be obtained from the analysis when items do not provide equal opportunities for different groups. The software in this analysis was SPSS 26.0 for validity testing. Meanwhile, the R studio program is used for IRT analysis, starting from model suitability testing, assumption testing, and DIF and reliability. The Standard Error of Measurement (SEM) uses the test of information function from the R Studio program. It is supported by Microsoft Excel software to convert the total score from theta (ability) from different strengths and levels of difficulty.

RESULTS AND DISCUSSION

The initial step in the analysis begins with testing the validity of the instrument using Exploratory Factor Analysis (EFA) with the help of SPSS software. It starts with the sample adequacy test using the KMO test (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) and Bartlett's Sphericity value, the Measure Sampling Adequacy (MSA) value. The analysis shows that the KMO value was 0.872. The Bartlett's Sphericity value was 0.000. And the MSA value for each item is more than 0.5. This means that the results of this analysis show that the sample met validity standards (Hair et al., 2010; Sarstedt & Mooi, 2014).

Furthermore, reliability estimates used an IRT. Reliability standards in IRT are seen from the information function value and Standard Error of Measurement (SEM). The total reliability result of this instrument is 0.922. Based on the marginal reliability output using the R studio program. The standard error of measurement (SEM) is seen in the Total Information Function (TIF) value.

Based on the information function value at capability (θ), 0.0 is 46.891 with a measurement error (SEM) of 0.146. These results indicate that the test produces optimal information when used on students with ability 0.0. Apart from this condition, there is also a strengthening of the condition regarding the Total Information Function (TIF) value, where if the TIF value is ≥ 10 then the test instrument is reliable for measuring students' numeracy skills. This analysis measures the strength of each item/question, which can explain the respondents' abilities as measured by the test (Myszkowski, 2021).

The valid and reliable instruments will continue to be analyzed at the next stage, namely gender bias analysis. Several stages of gender bias analysis were analyzed using Item Response Theory (IRT). This is an important step because the theory requires these stages and ensures gender bias analysis has gone through valid and reliable stages. These stages include: (1) the Model Fit Test; (2) the IRT Assumption Test; and (3) the DIF.

The model fit test was carried out to explain the characteristics of the items in the instrument. This test is seen from the Fit Indexes output results of the R studio program using the *irtGui* package (Yildiz, 2021). These results are presented in Table 1.

Table 1. Fit Indexes

Model	AIC	BIC	loglikelihood
Graded Response Model	17469.62	17852.58	-8637.81
GPCM	17375.55	17758.51	-8590.77

Table 1 shows the irtGui output results that provide two analysis recommendations: the Grade Response Model or GPCM. However, the final chosen was the GPCM model because

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

GPCM has the smallest AIC score. This is stated in the information in the irtGUI that the smaller the fit index of the model, the better the model (Desjardins & Bulut, 2018). GPCM is an analysis in which the questions are scored in tiered categories, but the difficulty index in each step is different/unordered. This means that in answering a question, the first step can be more difficult than the next step or vice versa, and GPCM only measures the level of difficulty and difference in power (Retnawati, 2014).

Next, the assumption test in IRT consists of the assumption test of unidimensionality, local independence, and parameter invariance. The unidimensional assumption test has been carried out simultaneously with construct validity using factor analysis (EFA), especially in the Total Variance Explained and Scree Plot table. Eigenvalue of 6.467 and Variance of 25.86% are the most dominant factors compared to other factors. The eigenvalue of the first factor is almost three times more than the second factor. Based on these results, the assumption test is fulfilled, following Mars's (2010) statement that the unidimensionality assumption is fulfilled if the first factor has an eigenvalue of at least two times comparing the second factor. In addition, the scree plot results also show the most dominant steepness indicated by the first factor of the seven factors. In sum, this instrument only measures one dimension, namely numeracy.

The next assumption test is the local independence test. This assumption defines that the response of the test respondent (Santri) to one test question does not affect the student's performance on other test questions. This assumption will be fulfilled if the student's response or answer to a test question does not influence the student's response to other test questions (Bahar et al., 2021; Retnawati, 2014).

Local independence has criteria for fulfilling a correlation value for each item lower than 0.20. In this study, the local independence test was carried out using Yen's Q3 (Chen & Thissen, 1997). The results of the local independence test showed the highest correlation value of 0.189, while the correlation values for the remaining items were lower than 0.200. Therefore, these findings indicate that the IRT assumption test criteria have been met.

Next, the last assumption is the parameter invariance. This assumption defines that the characteristics of test questions do not depend on the distribution of parameters of participants' skills tests. The parameters that characterize the respondent do not depend on the characteristics of the test questions. The implication is that test participants' skills will not change just because they respond to test questions with different levels of difficulty (Antara, 2020).

This assumption is verified by estimating the parameters of item and participant skills. Classification of even and odd items is used to measure item parameters. While participants' abilities are measured using the classification of odd and even respondents. With the help of the R Studio software program, it produces item parameter estimates and ability estimates based on differentiating power (a), level of difficulty, (b) pseudo guessing (c), and ability parameter estimates. The analysis results on this assumption also show that the item parameters and ability parameters based on parameters a, b, and c do not vary in the odd and even item groups. In other words, the parameter invariance assumption is met.

The next analysis is the model suitability test. This test aimed to explain the characteristics of the items in the instrument in more detail. This test is seen from the Fit Indexes output results of the R studio program using the irtGui package (Yildiz, 2021). The results are presented in Table 2.

Table 2. Fit Indexes

Model	AIC	BIC	loglikelihood
Graded Response Model	17469.62	17852.58	-8637.81
GPCM	17375.55	17758.51	-8590.77

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

Table 2 shows the irtGui output results that provide two analysis recommendations: the Grade Response Model or GPCM. However, the final chosen was the GPCM model due to the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BYC) having the smallest score. This is stated in the information in the irtGUI that the smaller the fit index of the model, the better the model (Desjardins & Bulut, 2018). GPCM is an analysis in which the questions are scored in tiered categories, but the difficulty index in each step is different/unordered. This means that in answering a question, the first step can be more difficult than the next step or vice versa, and GPCM only measures the level of difficulty and difference in power (Retnawati, 2014).

Next, item characteristics are based on item fit to verify that the IRT GPCM model was suitable for use because the items were declared fit. The fit provisions are based on the p.S_X2 value of more than 0.05. This means that if it is less than 0.05, the item does not fit the model. The results of the analysis are presented in Table 3.

Table 3. Item Fit

Question	S_X2	p, S_X2	Description
V1	54,18	1,18	Fit
V2	29,43	0,5	Fit
V3	15,63	0,93	Fit
V4	71,83	0,89	Fit
V5	63,18	0,95	Fit
V6	78,82	1,01	Fit
V7	30,65	0,49	Unfit
V8	39,39	0,86	Fit
V9	78,99	1,03	Fit
V10	32,65	0,34	Unfit
V11	36,96	0,91	Fit
V12	17,6	0,92	Fit
V13	19,51	0,77	Fit
V14	36,44	0,45	Unfit
V15	92,7	0,83	Fit
V16	77,03	0,74	Fit
V17	55,88	0,87	Fit
V18	43,44	1,05	Fit
V19	95,36	0,42	Unfit
V20	100,75	0,37	Unfit
V21	63,05	0,78	Fit
V22	131,08	0,97	Fit
V23	54,51	0,27	Unfit
V24	159,51	0,98	Fit
V25	185,87	0,92	Fit

The analysis results in Table 3 show that 19 items are fit and six are unfit. So, GPCM analysis is suitable to use because fit items are higher than unfit items.

Next, it analyzes the characteristics of each question item on the instrument using GPCM. This characteristic is taken from the difficulty level (b) and a different power (a). The provisions on the criteria are based on the differential power value which is in the logit scale of 0.00 – 2.00, and the level of difficulty in the logit scale is between -4.00 – 4.00 (DeMars, 2018). The results of the analysis are presented in Table 4.

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

Table 4. Characteristics Analysis of Numeracy Question Items

Items	Differential Power	Category	Difficulty Level	Category
Item_1	-0,182	Poor	-0,271	Good
Item_2	-0,083	Poor	-4,558	Very easy
Item_3	0,546	Good	3,591	Very Difficult
Item_4	-0,34	Poor	-0,375	Good
Item_5	-0,223	Poor	-11,131	Very easy
Item_6	0,328	Good	1,1674	Good
Item_7	-0,183	Poor	-4,87	Very easy
Item_8	0,944	Good	3,264	Very Difficult
Item_9	-0,546	Poor	-1,8632	Good
Item_10	0,656	Good	0,1	Good
Item_11	0,123	Good	1,654	Good
Item_12	-0,056	Poor	-20,473	Very easy
Item_13	1,348	Good	0,297	Good
Item_14	0,193	Good	5,852	Very Difficult
Item_15	-0,555	Poor	-1,5815	Good
Item_16	1,622	Good	0,7078	Good
Item_17	3,943	Poor	-0,11	Good
Item_18	6,8	Poor	-0,05	Good
Item_19	1,346	Good	0,294	Good
Item_20	-0,358	Poor	-1,5148	Good
Item_21	1,957	Good	0,225333	Good
Item_22	0,186	Good	3,374	Very Difficult
Item_23	1,998	Good	1,1648	Good
Item_24	-0,049	Poor	-4,6556	Very easy
Item_25	-0,04	Poor	-9,0266	Very easy
	Total	12	Total	16

The Table 4 shows the level of difficulty in this instrument is 64% in the good category, and only 48% has a good differential power. So, overall, this instrument meets the optimal level of difficulty and can differentiate between high- and low-achieving students. However, with these results, it might be said that this is the first indication of the existence/absence of differences between men and women.

The next analysis is to see the presence/absence of gender bias in the instrument using instrument analysis using DIF with the Mantel-Haenzel Method. The main reason for the analysis of gender factors is a general fact that often occurs in schools and in a statistical language, referred to as the irrelevance of the source of construct variance (Amelia et al., 2022). In this research, the role of gender analysis refers to identity, behavior, social order, and other lives that can influence one's perspective, either towards oneself or others, especially in Islamic boarding schools (Heidari et al., 2016). Item response theory using the DIF method can investigate this. At least, this investigation in an instrument setting is an initial basis for research regarding the presence/absence of gender bias in an environment (Wetzel et al., 2012). The results of the DIF analysis are presented in Table 5.

The criteria for determining DIF are, if the P value <0.05, then there is a piece of evidence that there is a significant difference in responding to items between groups. The results presented in Table 5 contain 1 question that is indicated as biased. This question discusses the connection between the battle of Badr and the beginning of fasting. This analysis can be a strengthening because this DIF analysis has explained the characteristics of items based on

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

bias, in this case, gender. For clarity, Figure 1 presents an overview of the DIF analysis results in diagram form.

Table 5. Analysis Result of DIF (Generalized Mantel-Haenszel chi-square statistic)

Number of Question	Stat.	P-value
Item 1	2.0270	0.3630
Item 2	3.3300	0.1892
Item 3	0.3602	0.8352
Item 4	1.3896	0.4992
Item 5	0.1213	0.9412
Item 6	0.9413	0.6246
Item 7	0.9629	0.6179
Item 8	4.3825	0.1118
Item 9	2.3544	0.3081
Item 10	2.2663	0.3220
Item 11	0.3436	0.8422
Item 12	0.6968	0.7058
Item 13	2.0700	0.3552
Item 14	0.5623	0.7549
Item 15	2.7150	0.2573
Item 16	0.7622	0.6831
Item 17	1.2263	0.5417
Item 18	0.3105	0.8562
Item 19	0.7334	0.6930
Item 20	3.3729	0.1852
Item 21	7.7652	0.0206 *
Item 22	0.7566	0.6850
Item 23	1.6453	0.4393
Item 24	2.1537	0.3407
Item 25	1.5040	0.4714

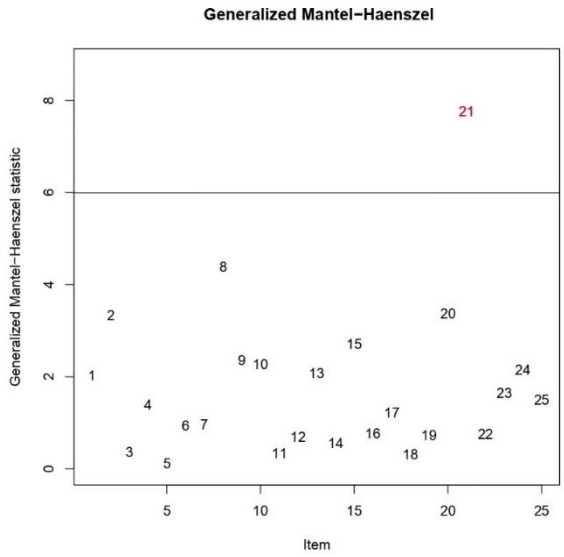

Figure 1. DIF Output Results

Figure 1 also shows that question number 21 indicates bias. Then, this question is deleted and not used for future tests as biased questions will have a negative effect on a certain group and vice versa. Question categories consist of appropriate, less appropriate, and inappropriate.

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

The FGD with Islamic boarding school teachers, question number 21, was deemed inappropriate, so it needs to be deleted or discarded. Also, this follows the opinion of Karkal & Kundapur (2016) and Odukoya et al. (2018), who state that inappropriate questions should not be used or deleted and not distributed to students.

In a clearer statement, the simple analysis in question number 21 is related to the understanding of Fiqh in mathematics teachers. The stimulus for this question is the incident of the Battle of Badr. One of the readings is about the incident on the day of the Battle of Badr, which occurred on Friday, the 17th of Ramadhan. The text of question no. 21 is "Based on the date of the Battle of Badr, what day is the 1st of Ramadhan 2 Hijriah?".

Male students are more difficult to answer. They think more about whether the beginning of the Hijriah is determined by reckoning or *rukyat*. In fact, the text clearly states that the 17th of Ramadan is the day of the battle of Badr. It means that the 1st of Ramadhan has passed. In contrast, in the next questions, which relate to the 1st of Eid al-Fitr, students can understand better because the word "*istikmal*" is mentioned.

The text and quantitative analysis show a gender bias for women. This may occur due to relating teaching to female students. The field of Fiqh, or the history of Islamic culture that is taught to women is still general. It means it is textual and more directed towards the implementation of Fiqh. For students who are in class 3 of Wustha or class IX Madrasah, female students are usually less concerned about phenomena such as the time of Eid compared to male students. Male students are more social and more critical of the phenomena and tend to blame for ignorance. Here, the role of male teachers is also fun to explain the differences between the time of Eid and enlightenment and to always respect differences.

These results indicate that item bias detection analysis is needed for mathematics teachers to identify that the question instrument has an appropriate construct. As a result, it functions to test two different groups, both men and women (Amelia et al., 2022). In the end, mathematics assessment instruments always relate to gender bias (Nathan & Umoinyang, 2022; Samritin, 2022). Moreover, the results of other studies also show that the gender variable influences mathematics ability, where specifically men are more interested and capable in physics and mathematics subjects, while women are more interested in biology (Steegh et al., 2019).

CONCLUSION

Based on the analysis, the instrument for testing students' numeracy abilities has met the validity and reliability values suitable for assessment use. The test instrument also has a suitable GPCM model and has met the IRT assumption test. Differential Item Functioning (DIF) with the Mantel-Haenzel method was used at the gender bias analysis stage. And 1 question was identified as gender bias. Based on these results, this instrument is suitable for use to meet the assessment standards to minimize gender bias in Islamic boarding schools.

ACKNOWLEDGMENT

The researcher would like to express his gratitude to the Directorate of Islamic Religious Higher Education (Diktis) of the Indonesian Ministry of Religion for fully funding this research through the Litapdimas 2023 program. The researcher would also like to thank all the leaders of Islamic boarding schools for their service and the students who are always ready and cooperative in following the numeracy test.

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

REFERENCES

Amelia, R. N., Astuti, S. R. D., & Sari, A. R. P. (2022). Deteksi bias gender pada instrumen evaluasi belajar kimia dengan metode Mantel-Haenzel. *JURNAL TARBIYAH*, *29*(2), 243. https://doi.org/10.30829/tar.v29i2.1781

Antara, A. A. P. (2020). *Penyetaraan vertikal dengan pendekatan klasik dan item response theory (teori dan aplikasi)*. Deepublish.

Ate, D., & Lede, Y. K. (2022). Analisis kemampuan siswa kelas VIII dalam menyelesaikan soal literasi numerasi. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *6*(1), 472–483. https://doi.org/10.31004/cendekia.v6i1.1041

Bahar, R., Istiyono, E., Widihastuti, W., Munadi, S., Nuryana, Z., & Fajaruddin, S. (2021). Analisis karakteristik soal ujian sekolah hasil musyawarah guru matematika di Tasikmalaya. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *10*(4), 2660. https://doi.org/10.24127/ajpm.v10i4.4359

Basharat, S. (2022). *Exploring gender gaps in mathematics achievement: The case of single-sex education in Saudi Arabia* [Universitetet i Oslo]. https://www.duo.uio.no/handle/10852/95343

Berndt, A. E. (2020). Sampling methods. *Journal of Human Lactation*, *36*(2), 224–226. https://doi.org/10.1177/0890334420906850

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289. https://doi.org/10.3102/10769986022003265

Clavel, J. G., & Flannery, D. (2023). Single-sex schooling, gender and educational performance: Evidence using PISA data. *British Educational Research Journal*, *49*(2), 248–265. https://doi.org/10.1002/berj.3841

DeMars, C. E. (2018). Classical test theory and item response theory. In *The Wiley Handbook of Psychometric Testing* (pp. 49–73). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118489772.ch2

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Etikan, I. (2017). Sampling and sampling methods. *Biometrics & Biostatistics International Journal*, *5*(6). https://doi.org/10.15406/bbij.2017.05.00149

Franklin, D., & Rangel, V. S. (2024). Estimating the effect of single-sex education on girls' mathematics and science achievement. *Leadership and Policy in Schools*, *23*(1), 97–114. https://doi.org/10.1080/15700763.2022.2108461

Gomez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benitez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104–110. https://doi.org/10.7334/psicothema2017.183

Hair, J. F., Black, W. C., & Babin, B. J. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Prentice Hall.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Heidari, S., Babor, T. F., De Castro, P., Tort, S., & Curno, M. (2016). Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. *Research Integrity and Peer Review*, *1*(1), 2. https://doi.org/10.1186/s41073-016-0007-6

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

Isbah, M. F. (2020). Pesantren in the changing Indonesian context: History and current developments. *QIJIS (Qudus International Journal of Islamic Studies)*, *8*(1), 65. https://doi.org/10.21043/qijis.v8i1.5629

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. https://doi.org/10.4103/2277-8632.191842

Khaerudin, K., & Munadi, S. (2020). *Pengembangan model evaluasi internal program pembelajaran pondok pesantren modern* [Universitas Negeri Yogyakarta]. https://eprints.uny.ac.id/70135/

Lee, P., & Joo, S.-H. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions*, *7*(1). https://doi.org/10.25035/pad.2021.01.004

Lee, S., & Kim, S. (2017). Detecting Differential Item Functioning based on gender: Field of mathematics in the TIMSS 2007. *Journal of Fishries and Marine Sciences Education*, *29*(3), 757–766. https://doi.org/10.13000/JFMSE.2017.29.3.757

Masqon, D. (2014). Dynamic of pondok pesantren as indegenous Islamic education centre in Indonesia. *EDUKASI: Jurnal Penelitian Pendidikan Agama Dan Keagamaan*, *12*(1). https://doi.org/10.32729/edukasi.v12i1.78

Peraturan Menteri Agama Republik Indonesia Nomor 31 Tahun 2020 tentang Pendidikan Pesantren, Pub. L. No. 31 (2020).

Montoya, S. (2018). *Defining literacy: UNESCO*. https://gamltest.uis.unesco.org/wp-content/uploads/sites/2/2018/12/4.6.1_07_4.6-defining-literacy.pdf

Muafiah, E., Sofiana, N. E., & Khasanah, U. (2022). Pesantren education in Indonesia: Efforts to create child-friendly pesantren. *Ulumuna*, *26*(2), 447–471. https://doi.org/10.20414/ujis.v26i2.558

Munthe, A. P. (2015). Pentingya evaluasi program di institusi pendidikan: Sebuah pengantar, pengertian, tujuan dan manfaat. *Scholaria : Jurnal Pendidikan Dan Kebudayaan*, *5*(2), 1. https://doi.org/10.24246/j.scholaria.2015.v5.i2.p1-14

Myszkowski, N. (2021). Development of the R library "jrt": Automated item response theory procedures for judgment data and their application with the consensual assessment technique. *Psychology of Aesthetics, Creativity, and the Arts*, *15*(3), 426–438. https://doi.org/10.1037/aca0000287

Nadal, K. L. (2017). Gender bias in education. In *The SAGE Encyclopedia of Psychology and Gender* (Vol. 01, pp. 37–47). SAGE Publications, Inc. https://doi.org/10.4135/9781483384269.n209

Nathan, N. A., & Umoinyang, E. I. (2022). Differential Item Functioning in Basic education certificate examination in Mathematics in Akwa Ibom State, Nigeria. *Journal of Research in Education and Society*, *13*(3), 136–146. https://doi.org/10.58579/AJB-SDR/4.1.2022.62

Odukoya, J. A., Adekeye, O., Igbinoba, A. O., & Afolabi, A. (2018). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Quality & Quantity*, *52*(3), 983–997. https://doi.org/10.1007/s11135-017-0499-2

OECD. (2019). Programme for international student assessment (PISA) results from PISA 2018.

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506

OECD Publishing, *III*, 1–10.

Rahman, M. M., Tabash, M. I., Salamzadeh, A., Abduli, S., & Rahaman, M. S. (2022). Sampling techniques (probability) for quantitative social science researchers: A conceptual guidelines with examples. *SEEU Review*, *17*(1), 42–51. https://doi.org/10.2478/seeur-2022-0023

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Ramdhani, S., Suryadi, D., & Prabawanto, S. (2021). Hambatan belajar matematika di pondok pesantren. *Jurnal Analisa*, *7*(1), 46–55. https://doi.org/10.15575/ja.v7i1.10106

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

Sahri, I. K., & Hidayah, L. (2020). Kesetaraan gender di Pesantren NU: Sebuah telaah atas single sex classroom di pendidikan diniyah formal Ulya Pondok Pesantren Al Fithrah Surabaya. *Journal of Nahdlatul Ulama Studies*, *1*(1), 67–105. https://doi.org/10.35672/jnus.v1i1.67-105

Samritin, S. (2022). Identifikasi muatan Differential Item Functioning pada data ujian nasional matematika. *Journal on Education*, *4*(4), 1675–1684. https://doi.org/10.31004/joe.v4i4.2508

Sarstedt, M., & Mooi, E. (2014). A concise guide to market research. In *A Concise Guide to Market Research*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-53965-7

Steegh, A. M., Höffler, T. N., Keller, M. M., & Parchmann, I. (2019). Gender differences in mathematics and science competitions: A systematic review. *Journal of Research in Science Teaching*, *56*(10), 1431–1460. https://doi.org/10.1002/tea.21580

Supriatna, D. (2018). Motivasi orang tua memilih pondokp pesantren untuk Anaknya. *Intizar*, *24*(1), 1–18. https://doi.org/10.19109/intizar.v24i1.1951

Wekke, I. S., & Hamid, S. (2013). Technology on language teaching and learning: A research on Indonesian Pesantren. *Procedia - Social and Behavioral Sciences*, *83*, 585–589. https://doi.org/10.1016/j.sbspro.2013.06.111

Wetzel, E., Hell, B., & Pässler, K. (2012). Comparison of different test construction strategies in the development of a gender fair interest inventory using verbs. *Journal of Career Assessment*, *20*(1), 88–104. https://doi.org/10.1177/1069072711417166

Whiteford, C. (2020). Mathematics, numeracy and literacy: A combination for success. *Practical Literacy*, *25*(2), 36–38. https://doi.org/10.3316/aeipt.228037

Yasid, A. (2018). *Paradigma baru pesantren: menuju pendidikan islam transformatif*. Yogyakarta: IRCiSOD.

Yildiz, H. (2021). IRTGUI: An R Package for Unidimensional Item Response Theory Analysis With a Graphical User Interface. *Applied Psychological Measurement*, *45*(7–8), 551–552. https://doi.org/10.1177/01466216211040532

Yusnita, E. (2011). Pembelajaran Kontekstual berlatar pondok pesantren pada materi garis dan sudut di kelas VII MTS. *Prosiding Seminar Nasional Penelitian, Pendidikan Dan Penerapan MIPA, Fakultas MIPA, Universitas Negeri Yogyakarta*, 11–18.

Copyright © 2024, Psychology, Evaluation, and Technology in Educational Research, ISSN 2622-5506