Al-Ishlah: Jurnal Pendidikan Islam

Volume 22 Issue 1, 2024
P-ISSN: 1693-7449, E-ISSN: 2685-6581
DOI: 10.35905/alishlah.v%vi%i.9649

Identification of HOTS-Based Islamic Education Tests in Junior High Schools : an Approach Bloom's Taxonomy and Rasch Model

Pipit Pitriani[1*], Ilzamudin Ma'mur[2], Enung Nugraha[3], Wasehudin[4] , Muhammad Habibi[5]

[1,2,3,4] *Postgraduate of Islamic Education Program, Universitas Islam Negeri Hasanuddin Banten, Indonesia*
[5] *Institute of Islam Hadhari, Universiti Kebangsaan Malaysia.*

**Corresponding E-mail:* 222621216.pipit@uinbanten.ac.id

Received date:	*Accepted date:*	*Published date:*
June, 12 2024	*June, 22 2024*	*June, 30 2024*

Abstract

Background: Higher-Level Thinking Skills (HOTS) are thinking abilities that need not only the ability to remember, but also the ability to think creatively and critically.
Research Objectives: This study identifies and analyzes the level of questions on the Islamic education Examination and analyzes the standardization of question items using the Rasch model technique.
Methods: Mixed method for analyzing document data with 160 multiple-choice questions. A quantitative approach was utilized to assess the quality of standard questions based on 31 students' accessible answer data. The document data was examined logically utilizing Bloom's Taxonomy Theory. Meanwhile, the data from students' answers was evaluated using the Item Response Theory (IRT) 1-parameter logistic model, also known as the Rasch model
Conclusions: The result shows that many PAI subject questions are still at the LOTS level. Meanwhile, in terms of the quality of the item test using the Rasch model approach, it was found that many items were still not fit, even though in general the reliability of the items was categorized as quite good.

Keywords: Islamic Education Test, HOTS, Rasch Model

Introduction

Thinking skills are the ability to carry out important thought processes in our daily lives. Thinking skills are an intellectual process that involves forming concepts through analysis, application, syntax, and evaluating information collected or produced by observation, experience, or reflection (Bhattacharya & Mohalik, 2021; Roets & Maritz, 2017). Thinking skills are closely related to a person's ability to use their cognitive and affective domains to obtain or provide information, solve problems, or make decisions in various active activities (Serevina et al., 2019; Wang & Wang, 2011). Therefore, thinking skills are a combination of cognitive processes and the ability to complete a given task (Ramos et al., 2013; Roets & Maritz, 2017).

Higher-Level Thinking Skills (HOTS) are thinking abilities that need not only the ability to remember, but also the ability to think creatively and critically. HOTS are the top level of cognitive processes. HOTS enables kids to overcome hurdles. Students who are frequently trained in HOTS can improve student achievement and reduce student weaknesses. High-level thinking skills that occupy the top level in the cognitive hierarchy of Bloom's Taxonomy are the

meaning of higher-order thinking skills (Hamdi et al., 2018). High-level thinking skills are complicated thinking processes in summarizing material, drawing conclusions, generating representations, analyzing, and building relationships that involve the most basic mental operations (Musliha et al., 2022).

Thinking skills are essential in the educational process. A person's thoughts can have an impact on their aptitude, quickness, and effectiveness during learning. As a result of the intimate association between thinking skills and learning, they are linked to the learning process (Resnick, 1987). Students who have been schooled in this manner of thinking typically have a good developmental impact on their schooling. Critical perception and processing are important stages in developing higher-order thinking skills (FitzPatrick & Schulz, 2015; Gil-Glazer et al., 2019).

The main goal of education is to produce individuals who think critically and creatively; this can be done by asking effective teachers in the classroom (Naufal & Alshaye, 2023). The form of questions asked must be able to stimulate students' minds to provide solution ideas to improve high-level thinking abilities. Students' abilities to develop high-level thinking skills must continue to be honed and trained through the learning process and tests containing HOTS question items (Hidayat et al., 2023; Kurniati et al., 2016; Mauji et al., 2020; Mustahdi, 2019).

The characteristics of HOTS questions contain complex elements, allow more than one solution, involve various decisions and interpretations, apply various criteria, and require various kinds of effort (Resnick,1987). The application of HOTS-based assessment is in line with learning that minimizes memory skills but must consider the skills of transferring one idea to another, processing and applying information, looking for relationships from various different pieces of information, solving problems using information, and critically reviewing concepts and information concepts. HOTS questions are applied as a tool to measure high-level thinking abilities, not just questions that are more difficult than memorizing questions. From a knowledge perspective, HOTS questions usually measure not only the factual, conceptual, and procedural dimensions but also the metacognitive dimension.

Several studies stated that students will understand a concept if they have high-level thinking skills (Amali et al., 2022; Awal Fikri Baharsyah, Muhammad Fauzi, Sabarudin, 2023; Gupta & Mishra, 2021; Sari et al., 2023; Virranmäki et al., 2020). Teachers must develop problems with high-level thinking skills to solve problems that exist in students' lives. Critical thinking is very necessary in the era of Revolution 4.0. Therefore, Islamic education or *Pendidikan Agama Islam* (PAI) teachers should change their way of thinking about the importance of exploring and stimulating students to think at a higher level.

HOTS-based assessment models have been widely developed and used in formal schools, beginning with the elementary, middle, and high levels (Muhajir & Hidayat, 2023).The development of an assessment model characterized by HOTS is also a priority for the Directorate of Islamic Education of the Ministry of Religion in conducting Islamic Religious

Education evaluations. Improving the quality of question instruments used in Islamic education assessments is an important task. This is consistent with efforts to develop students' competencies, which include not only understanding and knowing a type of knowledge or information, but also analytical, critical, and creative abilities in dealing with all problems that come their way, including questions. Teachers ask questions in class assessments, including tests in Islamic Education subjects. The aim is to determine the level of the cognitive domain in PAI subject evaluation questions. This is important to do as an evaluation for teachers and related parties to see the extent to which HOTS-based questions are implemented in accordance with the directions of the Ministry of Education and Culture and the Ministry of Religion so that they can catch up with other countries.

Therefore, HOTS-based PAI tests must also take into account the quality of standardization aspects in learning evaluation, such as discrimination, validation, item suitability, reliability, and distraction. Standardized tests, when developed and implemented properly, can be a useful tool in ensuring student success. Standardized testing is a good thing because it can create fairness for all students, ensure accountability in schools, and maintain high standards for educational reform.

One approach to measuring the quality of questions other than the classic test is to use the Rasch model approach. This model is one of the modern data analysis approaches to analyzing categorical data and overcomes various limitations of classical test theory. The Rasch model can be used as a technique for evaluating tests in schools. With the Rasch model, learning evaluation will be more objective because this model does not depend on sample characteristics, which tend to bias test results as in classical tests. Potential abilities such as talent or intelligence, as well as actual abilities, can also be measured. Through the features available in the Rasch model, various descriptions of the level of ability of students taking the test can be seen. One of the features available in the Rasch (Winstep) model is a map that combines the distribution of the abilities of students taking the test and the distribution of the difficulty levels of test items, which is not obtained with classical tests.

Studies related to HOTS analysis as well as analysis of the quality of HOTS questions in PAI subjects are still very limited (Sari et al., 2023), so specific studies are needed in terms of the object being analyzed, namely subject questions prepared by PAI teachers who joined in the Subject Teacher Deliberation or *Musyawarah Guru Mata Pelajaran* (MGMP). Previous studies were dominated by studies of subjects outside Islamic religious education. Apart from that, this paper also analyzes the quality of the questions using the Rasch model technique. With this study, it is hoped that it can provide input to PAI teachers who are part of the PAI MGMP to develop HOTS-based test instruments and question quality that pay attention to aspects of test standardization such as validity, reliability, level of difficulty, and test bias.

Methods

This study method is a mixed method. Qualitative methods are used to describe qualitative data obtained from the field (Creswell & Plano Clark, 2011; Hidayat et al., 2021). This research

was conducted in Tangerang Regency. The data obtained is in the form of questions on Islamic Religious Education (PAI) subjects for the last 4 years, namely the 2019/2020, 2020/2021, 2021/2022, and 2022/2023 academic years. The data obtained is documentary data and is available at the school. The number of questions available at school is 160 in multiple-choice form. For the analysis process, the researcher read and re-examined all the questions received, then examined each item using the Bloom's Taxonomy review table guide based on Bloom's cognitive theory. For review item categories based on Bloom's cognitive theory, they can be seen in Table 1.

Table 1. Review of Item Categories Based on Bloom's Cognitive

ITEM	LEVEL					
	LOTS	MOTS		HOTS		
	C1	C2	C3	C4	C5	C6
1						
2						
3						
4						
etc						

Indicator:
C1: Remembering, mentioning, defining
C2: Explain ideas/concepts
C3: Using information in different domains
C4: Determine aspects/elements
C5: make your own decisions
C6: Creating your own ideas

In defining the level of Bloom's taxonomy, the author's examination was then validated with two measurement and assessment specialists in the field of Islamic education subjects to avoid the author's bias and subjectivity. Meanwhile, to see the quality of PAI questions created by PAI teachers, item analysis was carried out using a quantitative approach and analyzed by Item Response Theory model 1 Parameter Logistics (PL) or what is usually called the Rasch Model (Ziniel, 2011).The aspects tested relate to the level of difficulty, suitability of the question items, and reliability of the questions. All item analysis uses Winstep Software.

Results and Discussion

The results obtained from the research have to be supported by sufficient data. The research results and the discovery must be the answers, or the research hypothesis stated previously in the introduction part.

1. Description of UAS Question Analysis

Table 2 shows an overview of PAI questions according to Bloom's taxonomy including the levels or levels included in Higher-Level Thinking Skills (HOTS).

Table 2. Bloom's Taxonomy Levels for PAI UAS Questions in the Last 4 Years

Cognitive Domain	School year			
Level	2019/2020	2020/2021	2021/2022	2022/2023

C1	15 (37.5%)	16 (40%)	15 (37.5%)	10 (25%)
C2	20 (50%)	13 (32.5%)	10 (25%)	12 (30%)
C3	5 (12.5%)	10 (25%)	13 (32.5%)	15 (37.5%)
C4	0 (0%)	1 (2.5%)	2 (5%)	3(7.5%)
C5	0 (0%)	0 (0%)	0 (0%)	0 (0%)
C6	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Data Source: Processed by Researchers

Based on Table 2, it shows that in the 2019/2020 academic year there was not a single HOTS-based question item. In the 2020/2021 academic year there is 1 HOTS-based PAI (2.5) question, and this has increased in 2021/2022 by 2 questions and in 2022/2023 by 3 questions. The HOTS level on the questions in the three academic years only reaches level C4.

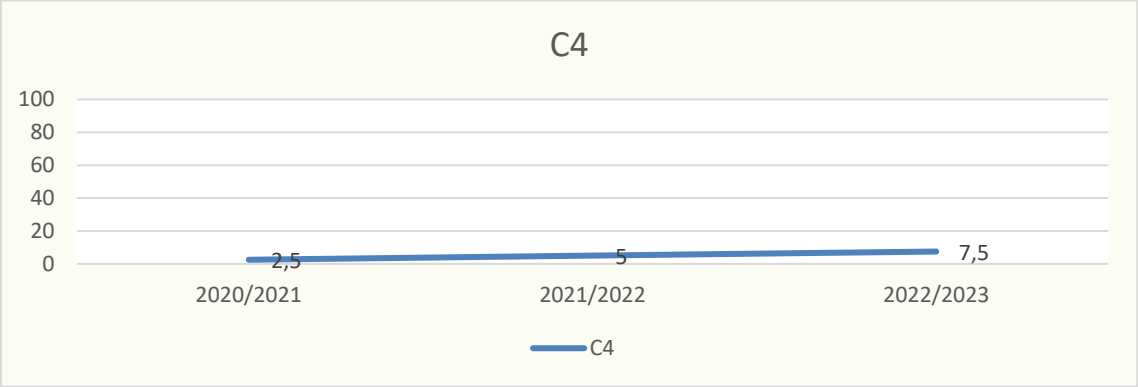


Figure 1. HOTS questions in the last 3 years

PAI UAS questions made by teachers are still dominated by questions Low Order Thinking Skills (LOTS), namely questions with levels C1 and C2. However, in the 2022/2023 academic year, Middle Order Thinking Skill (MOTS) questions account for 15 (37.5%) of the UAS questions.

In fact, to improve the quality of students' abilities, questions are needed that require high-level thinking abilities. The form of questions asked must be able to stimulate students' minds to provide solution ideas to improve high-level thinking abilities.

Nuro kehilangan uang sebesar Rp. 50.000,-. Suatu hari temannya menemukan uang tersebut ternyata lebih besar nominalnya dari uang Nuro yang hilang. Ketika uang tersebut diberikan kepada Nuro, Nuro menolaknya, perilaku Nuro tersebut menunjukkan sifat....

a. Jujur
b. Amanah
c. Istiqomah
d. Empati

Figure 2. Example of Level C3 Question Items

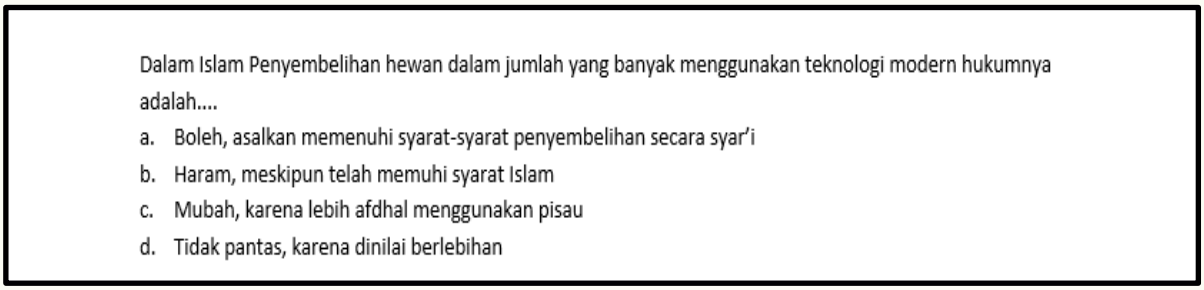
Figure 2 shows question level C3 (Application) because students need use knowledge in new situations. To reach this level, students need to remember and understand the concept of honesty.

Salat Jumat merupakan salat yang dilakukan pada hari Jumat dengan berjamaah, salat Jumat hukumnya fardu ain artinya wajib atas setiap muslim laki-laki yang sudah balig, dan berakal sehat, dengan peserta jamaahnya tidak kurang dari.....orang.

a. 3 orang
c. 30 orang

Figure 3. Example of Level C1 Question Items

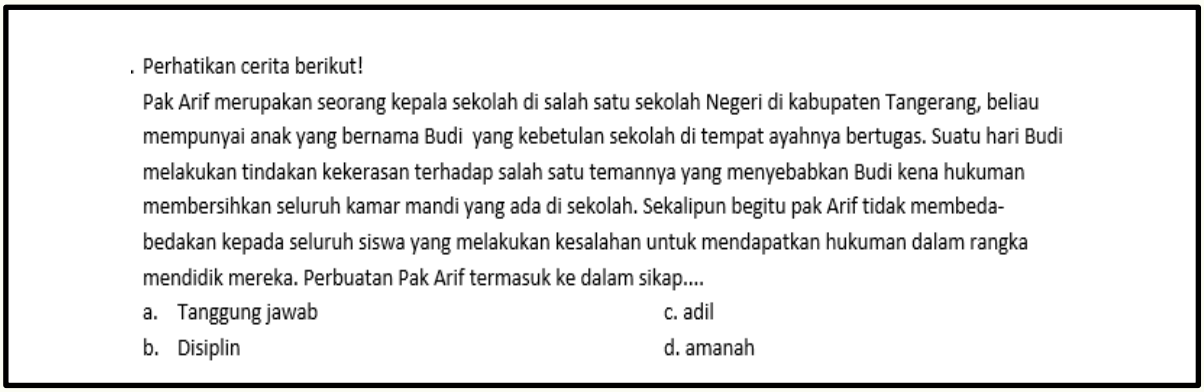
Figure 3 shows the questions at level C1 (remembering), students only need to remember or memorize those related to the requirements for congregational prayer, one of which is that there must be 40 people who are mature, have common sense and are experts in *mukim.*


Dalam Islam Penyembelihan hewan dalam jumlah yang banyak menggunakan teknologi modern hukumnya adalah....

a. Boleh, asalkan memenuhi syarat-syarat penyembelihan secara syar'i
b. Haram, meskipun telah memuhi syarat Islam
c. Mubah, karena lebih afdhal menggunakan pisau
d. Tidak pantas, karena dinilai berlebihan

Figure 4. Example of Level C2 Question Items

Figure 4 shows level C2 questions, namely understanding. Students are required to understand the concept of the requirements for animal slaughter. But before understanding, students need to memorize or remember the conditions for slaughtering animals.


. Perhatikan cerita berikut!

Pak Arif merupakan seorang kepala sekolah di salah satu sekolah Negeri di kabupaten Tangerang, beliau mempunyai anak yang bernama Budi yang kebetulan sekolah di tempat ayahnya bertugas. Suatu hari Budi melakukan tindakan kekerasan terhadap salah satu temannya yang menyebabkan Budi kena hukuman membersihkan seluruh kamar mandi yang ada di sekolah. Sekalipun begitu pak Arif tidak membeda-bedakan kepada seluruh siswa yang melakukan kesalahan untuk mendapatkan hukuman dalam rangka mendidik mereka. Perbuatan Pak Arif termasuk ke dalam sikap....

a. Tanggung jawab
b. Disiplin
c. adil
d. amanah

Figure 5. Example of Level C4 Question Items

Figure 5 shows an example of level C4 (analyzing) items. This level, students need to sort or describe the elements of the category of the concept of commendable actions.

2. Level difficulty Question Items

The level of difficulty of the questions is a description of the questions that fall into the difficult, medium or easy categories. The level of item difficulty can be explained by measuring items with logit units as depicted in Table 3 below.

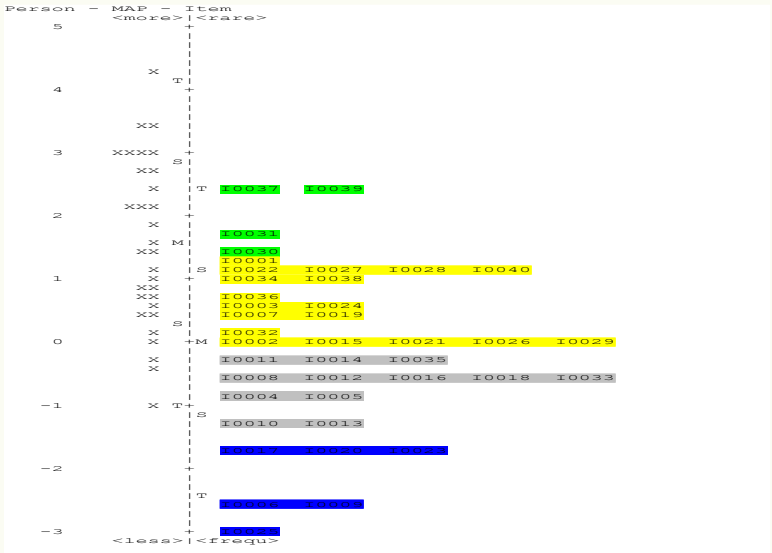


Figure 6. Question Item Map Item

Information:

	Very Difficult Question
	Difficult Problem
	Easy Question
	Very easy question

Based on Figure 6, it shows that there are 4 (10%) items questions that were categorized as very difficult, 18 (45%) of the questions were categorized as difficult, 12 (30%). The questions were categorized as easy and 6 (15%) of the questions were categorized as very easy. Items 37 and 39 have a categorized value of 2.38 logits most important questions difficult for students to answer. Meanwhile, question 25 has a value of-3.78 logit which is categorized as a question item the most answered correctly by students. Questions 37 and 39 are HOTS-based PAI questions.

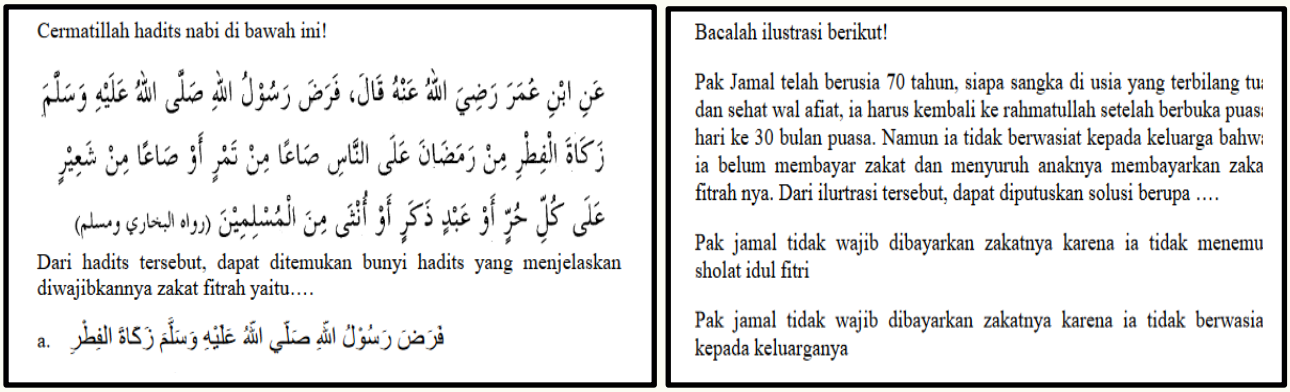


Figure 7. Question Items Categorized as HOTS

Questions 37 and 39 are categorized as HOTS because students need to have the ability to sort, differentiate, organize and connect each element. This shows that some students are still unable to answer HOTS-based questions.

Meanwhile, the questions that are categorized as the easiest are questions in category C1, as presented in Figure 8.

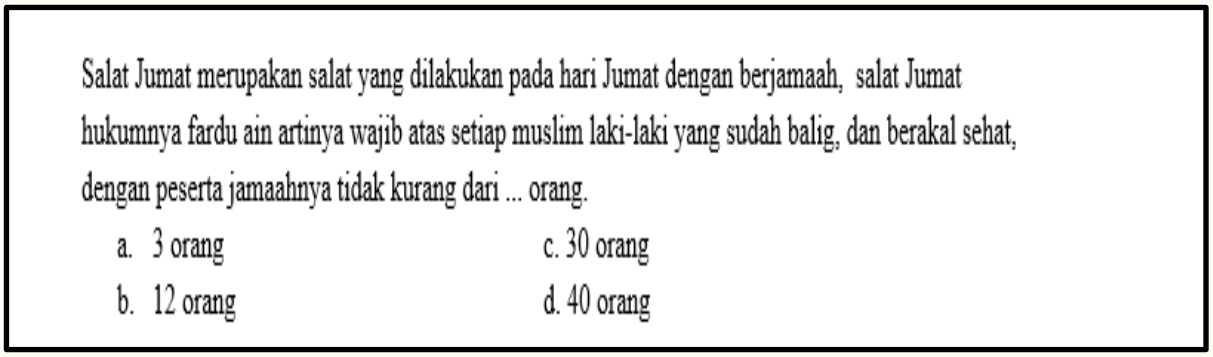


Figure 8. Question Items Categorized as HOTS

Figure 8 shows that item 25 is a C1 level item because it only emphasizes students' memorization or memory. Level C1 is categorized as the lowest level in Bloom's Taxonomy. There should not be too many questions at level C1 in the composition of exam questions. Furthermore, to determine the reliability of measurements, you can see the measurement function graph test that presented in Figure 9 the following.

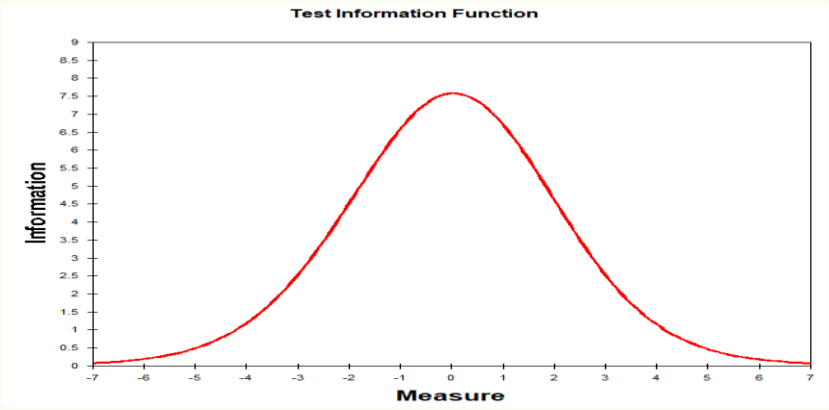


Figure 9. Test Function Information

Figure 9 show that the curve in the graph above is that the 40 questions given to 31 students show questions that are suitable for determining the abilities of only moderate students. However, the graph in the image shows the peak height is relatively high. This means that the reliability value is high. The higher the peak of the information function that can be achieved, the higher the measurement reliability (Bond & Fox, 2015; Boone et al., 2014).The evidence shown in Figure 4 is strengthened by the students' ability to answer the questions as shown in Figure 10.

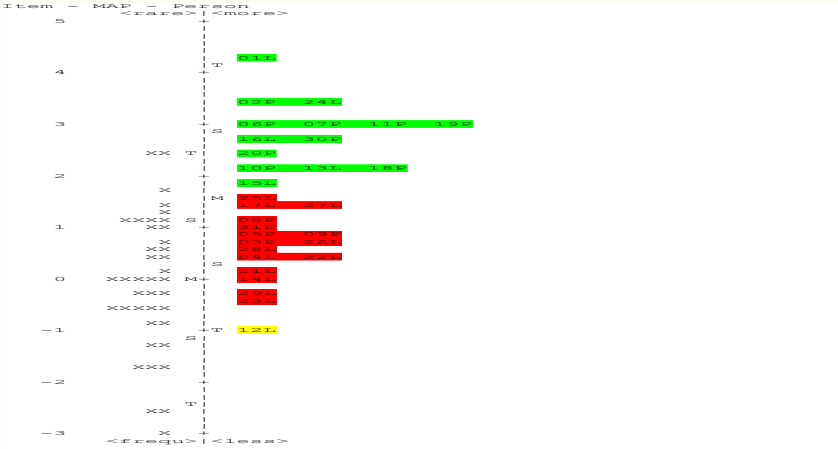


Figure 10. Person Map

Figure 10 shows that students tend to be more in the medium ability category, namely 16 (51.61%). Only 1 (3.22%) person was categorized as low ability level.

3. Conformity Question Items (Item Fit)

Information valuable other with Rasch modeling aside. The level of difficulty of an item is to see the quality of the item's suitability to the model, or what is abbreviated as item fit. Item fit works. It's normal to take measurements or not. If it is found that the question items do not fit then this is an indication that students have misconceptions about those items (Bond & Fox, 2015). Items fit is shown in Table 3 below.

Table 3. Criteria for Suitability of Question Items (Item Fit)

Items	MNSQ	ZSTD	PMC	Information
1	1.8691	2.5719	0.0600	Unfit
2	0.5396	-0.8295	0.5900	Fit
3	1.1092	0.4011	0.4000	Fit
4	0.3225	-0.7997	0.5900	Fit
5	0.7291	-0.0593	0.3600	Fit
6	3.4097	1.5434	-0.0800	Unfit
7	0.9613	0.051	0.4400	Fit
8	0.6159	-0.3694	0.4400	Fit
9	0.4141	-0.0696	0.2300	Unfit
10	0.4121	-0.3796	0.4500	Fit

Items	MNSQ	ZSTD	PMC	Information
11	0.9919	0.201	0.4000	Fit
12	0.4796	-0.6495	0.5400	Fit
13	0.4818	-0.2595	0.3900	Unfit
14	0.9944	0.201	0.4800	Fit
15	0.6186	-0.6194	0.5200	Fit
16	0.9883	0.2310	0.3300	Unfit
17	0.5216	-0.0095	0.3000	Unfit
18	1.1235	0.4111	0.2900	Unfit
19	1.2659	0.7113	0.4800	Fit
20	0.4958	-0.0395	0.3000	Unfit
21	0.5809	-0.7194	0.5400	Fit
22	1.0598	0.3011	0.5300	Fit
23	0.6331	0.1106	0.2800	Unfit
24	0.8205	-0.3592	0.4900	Fit
25	1,0000	0.0000	0.0000	Unfit
26	0.5757	-0.7294	0.5600	Fit
27	1.2873	0.9913	0.3600	Unfit
28	1.0224	0.171	0.4700	Fit
29	0.9392	0.0709	0.4200	Fit
30	0.6794	-1.2393	0.6600	Fit
31	0.6117	-1.5794	0.7200	Fit
32	1.3441	0.8013	0.2700	Unfit
33	1.4399	0.7714	0.1800	Unfit
34	0.9368	-0.0991	0.4800	Fit
35	2.1779	1.6222	-0.0700	Unfit
36	0.5168	-1.5695	0.7200	Fit
37	1.1920	0.6412	0.3800	Unfit
38	0.7950	-0.5792	0.5800	Fit
39	0.9208	-0.1191	0.5000	Fit
40	1.1084	0.4511	0.4400	Fit

Based on Table 3 show that of the 40 questions on the Final Semester Examination for the 2022/2023 Academic Year, there are 15 questions that are categorized as unfit or inappropriate. This shows that there are indications that there are misconceptions among students regarding these items. The question item is not appropriate because it does not comply with the criteria for the outfit means-squares value (MNSQ) received is 0.5 < MNSQ < 1.5, the Outfit Z-standard (ZSTD) value received is-2.0 < ZSTSD <+2.0 and the Point Measure Correlation (PMC) value is 0.4 < PMC < 0.85) (Boone et al., 2014).

4. Question Item Bias Detection

Something measurement valid is false one size. These questions do not contain bias. An instrument or question item is said to be biased if it is found that one individual with certain characteristics has an advantage over individuals with other characteristics. For example,

questions are easier for male students to answer than female students. To detect item bias, DIF (Differential Item Functioning) detection is used. An item question to be bias if found probability value of the item is below 5% (0.05). Bias detection the questions are shown in Figure 10 below.

Person CLASS	DIF MEASURE	DIF S.E.	Person CLASS	DIF MEASURE	DIF S.E.	DIF CONTRAST	JOINT S.E.	t	Welch d.f.	Welch Prob.	Mantel-Haenszel Chi-squ	Mantel-Haenszel Prob.	Size CUMLOR	Item Number	Item Name
L	-1.44	.80	P	.57	.70	-2.01	1.06	-1.90	28	.0681	.2353	.6276		16	I0016
L	2.70	.68	P	.02	.80	2.68	1.05	2.56	27	.0165	.2353	.6276	.41	30	I0030
L	1.90	.60	P	-2.09>	1.83	3.98	1.93	2.07	21	.0514	.0000	1.000		36	I0036
L	2.28	.63	P	2.48	.59	-.20	.87	-.24	28	.8157				37	I0037
L	1.22	.57	P	.57	.70	.65	.90	.72	27	.4801	.0000	1.000		38	I0038
L	2.70	.68	P	2.13	.59	.57	.90	.64	28	.5298	.2353	.6276		39	I0039
L	.90	.56	P	1.41	.61	-.52	.83	-.62	28	.5404	.2353	.6276		40	I0040

Figure 11. DIF Items

Item 30 is indicated to be biased because the probability value is <0.05. Meanwhile, the other 39 questions did not experience bias (p>0.05).

5. Reliability Item Questions and Person Test

Table 4 and Table 5 shows a summary of the reliability of items and individuals who took the Final School Examination (UAS) test for PAI subjects

Table 4. Summary of Measurements 400 Items

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	22.8	31.0	.00	.53	1.00	.1	.95	.0
S.D.	4.9	.0	1.19	.15	.21	.9	.56	.8
MAX.	30.0	31.0	2.38	1.05	1.66	3.0	3.41	2.6
MIN.	11.0	31.0	-2.52	.42	.64	-2.0	.32	-1.6
REAL RMSE	.57	TRUE SD	1.04	SEPARATION	1.81	Item	RELIABILITY	.77
MODEL RMSE	.55	TRUE SD	1.05	SEPARATION	1.89	Item	RELIABILITY	.78
S.E. OF Item MEAN = .19								

Measurements of the test instrument show that the reliability value of the questions is 0.77 with a model value of 0.78. This means that the value of 0.77 (Real RMSE) is in the medium range, namely 0.67 to 0.80. The value is 0.77 too can be concluded that the quality of the questions in the 2022/2023 PAI Final School Examination is categorized as having a fairly good reliability aspect.

Table 5. Summary of Measurements for 31 samples of test persons

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	29.7	40.0	1.58	.49	1.00	.1	.95	.1
S.D.	6.9	.0	1.29	.15	.14	.7	.56	.8
MAX.	39.0	40.0	4.25	1.04	1.27	1.5	3.68	2.8
MIN.	13.0	40.0	-1.03	.36	.76	-1.5	.19	-1.4

REAL RMSE	.52	TRUE SD	1.18	SEPARATION 2.27	Person RELIABILITY .84
MODEL RMSE	.51	TRUE SD	1.19	SEPARATION 2.33	Person RELIABILITY .84
S.E. OF Person MEAN = .24					

Person RAW SCORE-TO-MEASURE CORRELATION = .98
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .88

Table 5 shows that the test participant's reliability value is 0.84 (Real RMSE). This index value is in the good category (0.80 – 0.90).Sumintono & Widhiarso (2015) This also means that the consistency of answers from students sampled is categorized as good or good.

6. Unidimensionality

Unidimensionality the instrument is a measure which is important for measure what it should be be measured. Analysis Rasch model uses principal component analysis of the residual, i.e. measure the extent of diversity of the instrument measures what should. Table 6 shows the unidimensionality of the School Final Examination test instruments for PAI subjects for the 2022/2023 academic year.

Table 6: Unidimensionality (in Eigenvalue units)

	-- Empirical --		Modeled
Total raw variance in observations =	56.8 100.0%		100.0%
Raw variance explained by measures =	17.8 31.4%		31.8%
Raw variance explained by persons =	9.2 16.2%		16.4%
Raw Variance explained by items =	8.6 15.2%		15.4%
Raw unexplained variance (total) =	39.0 68.6%	100.0%	68.2%
Unexplned variance in 1st contrast =	4.0 7.0%	10.1%	
Unexplned variance in 2nd contrast =	3.9 6.8%	9.9%	
Unexplned variance in 3rd contrast =	3.4 6.1%	8.8%	
Unexplned variance in 4th contrast =	3.2 5.7%	8.2%	
Unexplned variance in 5th contrast =	2.9 5.1%	7.4%	

Based on Table 6 above, it shows that the raw data variance measurement results are 31.4%. This figure means that the requirement for unidimensionality of at least 20% has been met, if the value exceeds 40% it means it is better, especially if it is more than 60% it means special. Matter another, namely, the variance can not explained (unexplained *variance in contrast*) by the instrument should ideally not exceed 15% (Wu & Adams, 2007). All unexplained variance in contrast) shows values below 10%.

7. Answer Choices (Grading Scale)

Criteria for testing the suitability of answer choices on the PAI UAS test are used average observation value and Andrich Threshold value. Table 7 shows the validity of the answer choices.

Table 7. Validity of Answer Choices

Categories		Obsvd Avrge	Infit MNSQ	Outfit MNSQ	Andrich Thresholds
Label	Score				
1	A	0.23	1.11	1.17	NONE
2	B	0.40	0.97	0.98	-0.99
3	C	0.69	0.98	0.97	-0.69
4	D	0.90	0.86	0.88	0.11

The observed mean was found to increase consistently and evenly from 0.23 to 1.04 indicating uniformity of response patterns. Meanwhile, the Rasch-Andrich restriction moves from NONE ago to negative and then to positive sequentially (Baker, 2001). This shows that the scale answer choices are valid for the respondent.

Conclusion

This study shows that in general the questions for the Middle School PAI Subject Exams made by MGMP over the last 4 years are still at low and medium ability levels. Only a few questions are HOTS based. As a result, PAI teachers should prepare more HOTS-based questions in the hopes that they will increase the quality of students' thinking at a higher level. The quality of the PAI Final Exam questions for the 2022/2023 Academic Year is still categorized as unfit or inappropriate. This shows that there are indications happen students' misconceptions about these items. PAI teachers improve the quality of their teaching so that misconceptions can be avoided when teaching it again

Acknowledgments (Optional)

We would like to thank all the participants for volunteering their time to participate in this study.

References

Amali, L. N., Anggani Linggar Bharati, D., & Rozi, F. (2022). The Implementation of High Order Thinking Skills (HOTS) Assessment to Evaluate the Students' Reading Comprehension Achievement. *English Education Journal*, *12*(1), 10–18. https://doi.org/10.15294/eej.v12i1.52571

Awal Fikri Baharsyah, Muhammad Fauzi, Sabarudin, Y. S. (2023). Kemampuan Mahasiswa dalam Menyelesaikan Soal-Soal HOTS pada Materi PAI Sekolah/Madrasah. *Jurnal Tawadhu*, *7*(1), 1–23.

Baker, F. B. (2001). *The Basics of Item Response Theory* (Second). ERIC Clearinghouse on Assessment and Evaluation.

Bhattacharya, D., & Mohalik, R. (2021). *Factors Influencing Students ' Higher Order Thinking Skills Development*. *March*.

Bond, T. ., & Fox, C. . (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research*. Sage Publications, Inc.

FitzPatrick, B., & Schulz, H. (2015). Do Curriculum Outcomes and Assessment Activities in

Science Encourage Higher Order Thinking? *Canadian Journal of Science, Mathematics and Technology Education*, *15*(2), 136–154. https://doi.org/10.1080/14926156.2015.1014074

Gil-Glazer, Y., Walter, O., & Eilam, B. (2019). PhotoLingo—Development and Improvement of Higher-Order Thinking and Language Skills Through Photographs. *Journal of Education*, *199*(1), 45–56. https://doi.org/10.1177/0022057419843523

Gupta, T., & Mishra, L. (2021). Higher-Order Thinking Skills in Shaping the Future of Students. *Psychology and Education*, *58*(2), 9305–9311.

Hamdi, S., Suganda, I. A., & Hayati, N. (2018). Developing higher-order thinking skill (HOTS) test instrument using Lombok local cultures as contexts for junior secondary school mathematics. *Research and Evaluation in Education*, *4*(2), 126–135. https://doi.org/10.21831/reid.v4i2.22089

Hidayat, W., Anzali, M. N., & Turmudi, M. (2023). Speaking English Performance Assessment with the Facet Rasch Measurement Model. *Jurnal Evaluasi Pendidikan*, *14*(1), 8–11. https://doi.org/10.21009/jep.v14i1.38987

Hidayat, W., Musab, M., Lawahid, N. A., & Mujahidah, M. (2021). Developing the flipped learning instrument in an ESL context: The experts' perspective. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *25*(1), 35–48. https://doi.org/10.21831/pep.v25i1.38060

Kurniati, D., Harimukti, R., & Jamil, N. A. (2016). Kemampuan berpikir tingkat tinggi siswa SMP di Kabupaten Jember dalam menyelesaikan soal berstandar PISA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 142–155. https://doi.org/10.21831/pep.v20i2.8058

Mauji, S. M., Mulyanti, Y., & Nurcahyono, N. A. (2020). Analisis Kesalahan Siswa dalam Menyelesaikan Soal Trigonometri Berdasarkan Teori Newman. *De Fermat : Jurnal Pendidikan Matematika*, *2*(2), 77–82. https://doi.org/10.36277/defermat.v2i2.44

Muhajir, M., & Hidayat, W. (2023). Fıqh Subject Exam Questıons Analysıs : Is ıt Based on HOTS ? *Tuijin Jishu/Journal of Propulsion Technology*, *44*(5), 868–884. https://doi.org/10.52783/tjjpt.v44.i5.2706

Musliha, S., Sudana, D., & Wirza, Y. (2022). The Analysis of Higher Order Thinking Skills (HOTs) in the Test Questions Constructed by English Teachers. *Proceedings of the Fifth International Conference on Language, Literature, Culture, and Education (ICOLLITE 2021)*, *595*(Icollite), 610–617. https://doi.org/10.2991/assehr.k.211119.095

Mustahdi. (2019). *Modul Penyusunan Soal Kemampuan Berpikir Tingkat Tinggi (HOTS) Mata Pelajaran PAI dan Budi Pekerti*. Direktorat Pembinaan Sekolah Menengah Atas, Direktorat Jenderal Pendidikan Dasar Dan Menengah, Kementerian Pendidikan Dan Kebudayaan.

Naufal, M. A., & Alshaye, I. A. (2023). The Effectiveness of a Realistic Mathematics Education to Increase Higher Order Thinking Skill (HOTS) of Secondary School Students. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *7*(1), 54–59. https://doi.org/10.21831/pep.v20i2.8058

Ramos, J. L. S., Dolipas, B. B., & Villamore, B. B. (2013). Higher Order Thinking Skill and Academic Performance in Physics of College Students: A Regression Analysis. *International Journal of Innovative Interdisciplinary Research*, *4*, 48–60.

Resnick, L. B. (1987). *Education and learning to think*. National Academy Press.

Roets, L., & Maritz, J. (2017). Facilitating the development of higher-order thinking skills (HOTS) of novice nursing postgraduates in Africa. *Nurse Education Today*, *49*, 51–56. https://doi.org/10.1016/j.nedt.2016.11.005

Sari, I., Usama, D., Noviani, D., & Basuni, F. (2023). Langkah Penyusunan dan Analisis Butir Soal Hots (Higher Order Thinking Skills) pada Mata Pelajaran Pendidikan Agama Islam (PAI)

dan Budi Pekerti. *Jurnal Insan Pendidikan Dan Sosial Humaniora*, *1*(4), 56–73. https://doi.org/10.59581/jipsoshum-widyakarya.v1i4.1605

Serevina, V., Sari, Y. P., & Maynastiti, D. (2019). Developing high order thinking skills (HOTS) assessment instrument for fluid static at senior high school. *Journal of Physics: Conference Series*, *1185*(1), 0–9. https://doi.org/10.1088/1742-6596/1185/1/012034

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Trim Komunikata Publishing House.

Virranmäki, E., Valta-Hulkkonen, K., & Pellikka, A. (2020). Geography tests in the Finnish Matriculation Examination in paper and digital forms – An analysis of questions based on revised Bloom's taxonomy. *Studies in Educational Evaluation*, *66*(May), 100896. https://doi.org/10.1016/j.stueduc.2020.100896

Wang, S., & Wang, H. (2011). Teaching Higher Order Thinking in the Introductory MIS Course: A Model-Direct Approach. *Journal Education for Business*, *86*, 208–212. https://doi.org/10.1080/08832323.2010.505254

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions.

Ziniel, W. (2011). *Third Party Product Reviews and Consumer Behaviour: A Dichotomous Measuring via Rasch, Paired Comparison and Graphical Chain Models*. Springer Gabler.