ELEMENTARY *Islamic Teacher Journal*
E-ISSN : 2503-0256 / ISSN : 2355-0155
Volume 12 Number 1 January - June 2024 (PP. 117-134)
http://dx.doi.org/10.21043/elementary.v12i1.27841
Diakses di : http://journal.iainkudus.ac.id/index.php/elementary

Exploration The Characteristics of Indonesian Final Semester Assessment Questions Using The Rasch Model

Rahmat Danni*[1], Banun Havifah Cahyo Khosiyono[2], Berliana Henu Cahyani[3] Amir Hamzah[4], Rezkilaturahmi[5]
[1,2,3]Universitas Sarjanawiyata Tamansiswa, Yogyakarta, Indonesia
[4]Universitas Islam Negeri Raden Fatah, Indonesia
[4]Universitas Negeri Yogyakarta, Indonesia
**corresponding author: rahmatdani93@gmail.com*

Abstract

An effective and valid end-of-semester assessment is crucial for accurately measuring students' abilities in Indonesian language subjects. This study aims to describe the characteristics of end-of-semester assessment items for Indonesian subjects using the Rasch model. This research is a quantitative descriptive study involving 36 students from grade V at Islamic Elementary School Private (MIN) 1 East Belitung. The data consisted of responses to 30 multiple-choice questions from the end-of-semester assessment. The results indicate that the end-of-semester assessment instrument for Indonesian for grade V at MIN 1 East Belitung for the 2023/2024 academic year meets the criteria for a good test instrument based on the Rasch model approach, with 77% of the items being fit and 23% not fit. The instrument's reliability coefficient item reliability at 0.77, person reliability at 0.80, and Cronbach's alpha at 0.84 demonstrate consistency and reliability in measuring student abilities. Items that are not fit (numbers 3, 6, 7, 11, 12, 19, and 23) are unsuitable for assessing student abilities and need to be revised or removed. The use of non-fit items may result in invalid information. This study is limited to multiple-choice questions; therefore, further research should analyze items in descriptive or mixed multiple-choice and descriptive formats using the Rasch model approach. The impact of this study is expected to contribute to the improvement and development of assessment instruments, thereby enhancing the accuracy of student evaluations.

Keywords: *Indonesian Language, Islamic Elementary School, Islamic Elementary School Private, Rasch Model*

INTRODUCTION

In this disruptive era, educators are required to train and maximize students' abilities, including both hard skills and soft skills. A well-structured learning system is crucial to optimizing students' abilities. (Mardapi, 2016) emphasizes that a good learning system can produce quality learning outcomes.

This perspective is supported by the current trend of integrating student-centered learning models, which enable students to explore and develop their own abilities more effectively. According to Akyol & Garrison (2011), student-centered learning promotes the development of 21st-century skills, particularly higher-order thinking skills, which are essential for students' future success.

In addition, besides having a good learning system, an educator also needs to build an appropriate and objective assessment system. An appropriate assessment system can improve students' abilities Istiyono, Mardapi, & Suparno (2014). Meanwhile, an objective assessment system can produce accurate information regarding the quality and success of learning (Anderson & Krathwohl (2001). In other words, the success of the learning process carried out by educators can be known accurately through learning outcomes measured using a good assessment system.

However, assessment is collecting information and interpreting student achievements (Stiggins & Chappuis, 2012). Danni, Wahyuni, & Tauratiya (2021) define assessment as the process of interpreting measurement results. Thus, an assessment can be made if there is a measurement. Meanwhile, measurements can be carried out if there are good instruments or measuring tools. Assessments carried out based on measurement results with inappropriate instruments can produce wrong information (de la Torre & Minchen, 2014; Istiqlal et al., 2024; Knorn, Topalovic, & Varagnolo, 2022). Therefore, instruments are needed that meet good criteria so that the information obtained is accurate so that it can be used as a basis for making decisions or policies.

Furthermore, Reynolds, Livingston, & Willson (2009) stated that an instrument can be said to be of quality if the validity and reliability criteria are met. It was also emphasized by Sarea & Ruslan (2019) that the criteria for validity and reliability are the main requirements for an instrument to be considered quality. Instruments that meet the validity criteria are proven by the instrument's accuracy in measuring student competence (Ramadhan, Mardapi, Prasetyo, & Utomo, 2019). Meanwhile, an instrument is said to be reliable if the instrument is tested many times to produce consistent/steady measurement results (Sudjana, 2016). Thus, a quality instrument must be accurate and consistent when tested or used to measure students' abilities.

Apart from the validity and reliability criteria, there are other criteria so that an instrument can be said to be of quality, namely the criteria for level of difficulty, distinguishing power, and distracting power (Ahsani & Taqiyah, 2024; Ramadhan, Sumiharsono, Mardapi, & Prasetyo, 2020). These criteria are popular in the classical test theory approach (Arlinwibowo, Retnawati, & Kartowagiran, 2021). The classical test theory approach generally assumes that the observed score is an accumulation of pure scores and errors ($O_i = T_i + E_i$) (Himelfarb, 2019). The classical test theory approach is still used in the learning assessment system in Indonesia. It was later discovered that there was a weakness in the classical test theory approach, namely that the measurement score depended on the characteristics of the items being tested and the characteristics of the items depended on the test taker's abilities (Retnawati, 2014). Therefore, a new approach emerged namely item response theory, as an answer to the weaknesses of classical test theory.

In addition, if in the classical test theory approach, there are three criteria, then in item response theory there are three logistic parameter models (PL), namely level of difficulty (b_i), distinguishing power (a_i), and pseudo-guess (c_i). There is also a simpler model known as the Rasch model. The Rasch model is a measurement model that is in line with item response theory but is simpler because it only has one parameter, namely the level of difficulty (b_i) (Fajrianthi, Hendriani, & Septarini, 2016). The measurement results using the item response theory approach do not depend on the characteristics of the questions being tested but are obtained from the chances of the test taker answering the questions correctly. This is reflected in the assumptions that must be met in the item response theory approach, namely unidimensionality, local independence, and parameter invariance (DeMars, 2018).

The unidimensional assumption indicates that the questions only measure one ability (Arlinwibowo et al., 2021). The local independence assumption aims to prove that test item scores do not depend on other items (Zeigenfuse, Batchelder, & Steyvers, 2020). DeMars (2018) states that the local independence assumption will be met if the unidimensional assumption is met. Meanwhile, the parameter invariance assumption requires that the item parameters do not depend on the test taker and vice versa (Abdellatif, 2023). By fulfilling these assumptions, it is a guarantee that the measurement results using the item response theory approach are more accurate than classical test theory. Students with high ability have a chance of being able to answer difficult questions and conversely, students with low ability have a small chance of answering difficult items. Therefore, a system for assessing student learning outcomes needs to be developed using an item response theory approach.

Therefore, in contrast to classical test theory, item response theory is currently not widely used in student learning outcomes assessment systems, especially at the basic education level (Wilson, 2023). One of the schools that applies the classical test theory assessment system is MIN 1 East Belitung, especially in Indonesian language subjects (Rezi, 2024). Meanwhile, to find out the quality of the teacher's question items, it is not through testing the questions, but by peer assessment or scientific knowledge. In this way, the quality of the questions developed to measure students' abilities is unknown because they have not been empirically tested. However, measuring students' abilities must use quality instruments so that the measurement results are accurate and objective.

Based on the problems that have been described, it is necessary to analyze the questions so that the quality of the questions developed by teachers at MIN 1 East Belitung can be determined empirically. Therefore, this research aims to analyze and describe the quality of Final Assessment Questions (PAS) in Indonesian subjects at MIN 1 East Belitung. The quality of the test items is described through validity verification, reliability estimates, and item characteristics using the Rasch model.

METHODS

This research is classified as descriptive research with a quantitative approach. A quantitative approach is utilized in this research to provide a precise and objective assessment of the quality of the PAS questions. This approach allows for the systematic evaluation of item parameters and provides a clear, numerical representation of question quality, which is essential for identifying specific areas for improvement and ensuring the reliability and validity of the assessment tools based on the Rasch model. The aim of this research is to describe the quality of odd semester (PAS) questions for the 2023/2024 academic year in Indonesian subjects at MI/MIN 1 East Belitung using the Rasch modeling approach. The research data consists of secondary data, specifically the responses from all grade V students of MIN 1 East Belitung on the PAS question for the Indonesian subject, totaling 36 students. The questions tested were 30 items in the form of multiple choices with 4 alternative answers. The data analysis technique uses the Rasch model to determine the characteristics of the questions in the form of validity, reliability, and item difficulty (b_i) parameters which are analyzed with the help of R and Winstep software. In addition, a question item can be declared feasible/acceptable if it has an Outfit mean square (MNSQ) value in the range 0.5<MNSQ<1.5, Outfit Z-standard (ZTSD) in the range -2 <ZSTD<+2, and point measure correlation in the range 0, 3<PMC<0.85 and the difficulty level parameter index (b_i) in the range of 0 to +2 (Bond & Fox, 2015; Boone, Staver, & Yale, 2014; Fauziana & Wulansari, 2021; Mustafa, Khairani, & Ishak, 2021; Sumintono & Widhiarso, 2015) and

Reliability values above 0.70 indicate good reliability (Cronbach, 1951; Tavakol & Dennick, 2011). Therefore, if there are questions that do not meet these criteria, they are declared invalid or not suitable for use.

RESULTS AND DISCUSSION

Unidimensionality and Local Independence Assumptions

The first assumption that needed to be met in analyzing questions using the Rasch model was the unidimensional assumption. The unidimensional assumption aimed to find out that the items only measure one dimension (Edelen & Reeve, 2007). In other words, the dimension measured in this research instrument was the Indonesian ability of class V students at MIN 1 East Belitung. Proving the unidimensional assumption was carried out using exploratory factor analysis techniques. Students' responses to the PAS questions of the Indonesian subject were analyzed using exploratory factor analysis to examine the assumption of unidimensionality. The unidimensional assumption could be known through eigenvalues or scree plots. The analysis results were displayed in figure 1

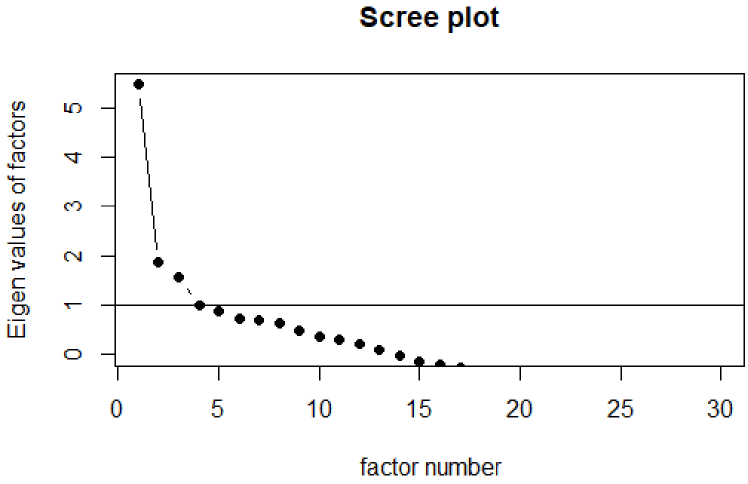


Figure 1. Scree plot of eigenvalues for Indonesian language questions

Based on figure 1, it could be explained that the X-axis (Factor Number) was the factor/dimensional number formed from the Indonesian PAS questions. These factors were ordered by the amount of variance explained, from largest to smallest. The Y axis *(Eigenvalues of Factors)* showed the eigenvalues of each factor. Eigenvalues were a measure of the amount of variance explained by the factor and points *(Eigenvalues)* represent the eigenvalues for one factor (Gana & Guillaume Broc, 2019).

In addition, in figure 1, it could be seen that there was 1 point that was highest compared to the other points. This showed that there was one dominant factor/dimension in the Indonesian PAS instrument. The steepness of the scree plot indicated the number of dimensions formed, while the sloping shape indicated that there were no dimensions formed (Susetyo, 2015). Thus, based on the *scree plot* formed, it could be concluded that there was one dimension measured by the Indonesian PAS instrument. Apart from using *a scree plot*, the unidimensional assumption could also be proven through eigenvalues. The eigenvalues of the Indonesian PAS questions were presented in Table 1.

Table 1. Eigenvalues of Indonesian PAS questions

Factor	Total	% variant	Cumulative %
1	6,224	20,746	20,746
2	2,676	8,920	29,665
3	2,566	8,553	38,218
4	1,800	5,999	44,217
5	1,718	5,728	49,945

showed the 5 highest eigenvalues of the Indonesian PAS questions at MIN 1 East Belitung. Factor 1 formed had the highest eigenvalue compared to other factors, namely 6.224 with a cumulative percentage of 20.746%. When compared with factor 2 and factor 3, the eigenvalue of factor 1 had a large proportion and the difference in value was 2 times compared to factors 2 and 3. The difference in eigenvalue was many times greater between factor 1 and the other factors, indicating that there was 1 dominant factor being measured (Gana & Guillaume Broc, 2019). Based on the scree plot and eigenvalues, it could be concluded that the Indonesian PAS question instrument measures one dimension so that it met the unidimensional assumption. Unidimensionality meant that a set of questions or items in a test or instrument measures a single underlying trait or construct (Xu & Stone, 2012). This was important in psychological or educational measurement to ensure that the test was focused on a specific attribute. In this case, the instrument measures the Indonesian proficiency of grade V students at the MIN 1 East Belitung.

The assumption of local independence will be fulfilled if the instrument was unidimensional (DeMars, 2018; Retnawati, 2016). Thus, the fulfillment of the unidimensional assumption in the Indonesian PAS question instrument proved that the question items did not depend on other items so that the assumption of local independence was also fulfilled.

Parameter Invariance Assumption

The assumption of parameter invariance could be known by estimating item parameters by dividing them into two odd and even groups and then forming a scree plot (Danni et al., 2021). The parameter invariance assumption was fulfilled if the parameter distribution approached a diagonal line (Abdellatif, 2023; Tran, Dorofeeva, & Loskutova, 2018). The results of the parameter invariance assumption test on the Indonesian PAS questions were presented in Figure 2 and Figure 3.

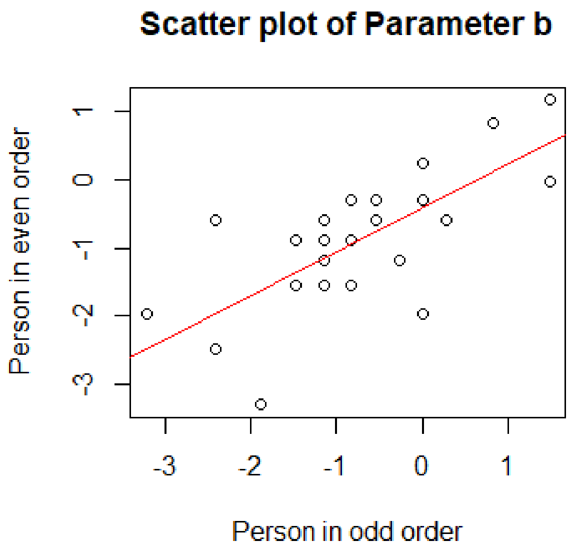


Figure 2. Distribution of difficulty level parameters (b_i)

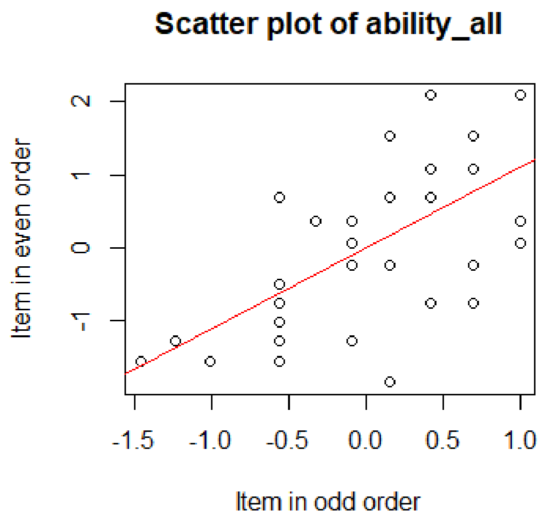


Figure 3. Distribution of student abilities for odd and even numbers (θ)

Figure 2 showed that the distribution of difficulty level parameters for the PAS questions in Indonesian subjects at MIN 1 East Belitung questions was close to the diagonal line and Figure 3 also showed that the distribution of student ability parameters was close to the diagonal line. In this way, the assumption of invariance of the parameters of the Indonesian PAS questions was fulfilled so that the analysis of the questions using the Rasch model could be continued.

Characteristics of Indonesian PAS questions based on the Rasch model

Items analyzed using the Rasch model approach can provide information regarding the characteristics of the items such as item difficulty level, instrument reliability, and information function, while item validity could be determined based on the extent to which the items met the criteria for good/fit items in the Rasch model approach. Criteria for good/fit items in the Rasch model approach included having an Outfit Mean Square (MNSQ) value in the range 0.5<MNSQ<1.5, Outfit Z-standard (ZTSD) in the range -2 <ZSTD<+2, and point measure correlation in the range 0.3<PMC<0.85 and the difficulty level parameter index (bi) in the range 0 to +2 (Bond & Fox, 2015; Boone et al., 2014; Danni & Tauratiya, 2020; Fauziana & Wulansari, 2021; Mustafa et al., 2021; Sumintono & Widhiarso, 2015). The results of the analysis of PAS Indonesian questions using the Rasch model approach with the help of Winstep software were presented in Table 2.

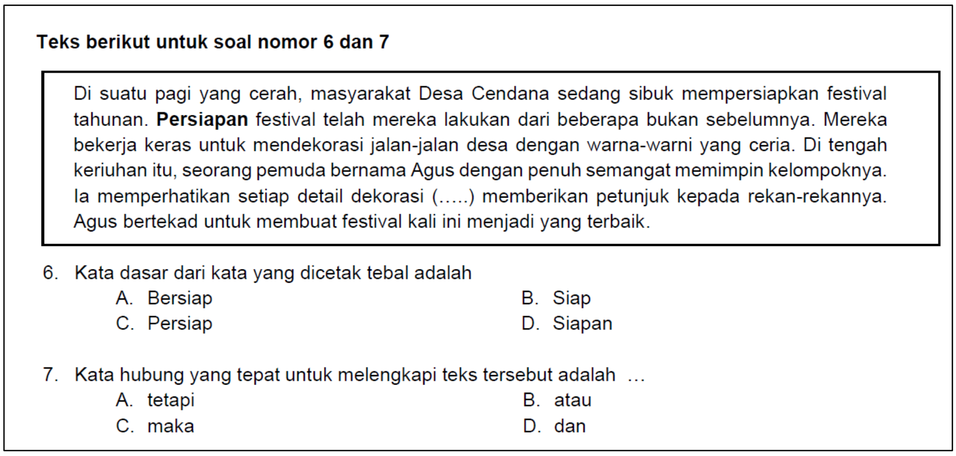
Table 2. Output characteristics of Indonesian PAS questions

Item	Total score	b_i	Infit		Outfit		PMC	Category
			MNSQ	ZSTD	MNSQ	ZSTD		
B1	19	0.73	0.84	-1.1	0.76	-1	0.58	Fit
B2	24	0	0.88	-0.7	0.86	-0.3	0.51	Fit
B3	32	-1.62	1.12	0.4	**2.24**	1.4	**0.12**	Unfit
B4	24	0	1.16	0.9	1.11	0.4	0.31	Fit
B5	32	-1.62	0.81	-0.4	0.58	-0.3	0.42	Fit
B6	27	-0.49	1.29	1.4	1.36	0.9	**0.16**	Unfit
B7	27	-0.49	1.13	0.7	**2.89**	**3**	**0.18**	Unfit
B8	22	0.3	1.03	0.3	1	0.1	0.41	Fit
B9	28	-0.68	0.92	-0.3	0.81	-0.2	0.43	Fit
B10	13	1.6	0.94	-0.3	0.81	-0.7	0.52	Fit
B11	24	0	1.47	2.5	**2.71**	3.6	**-0.04**	Unfit
B12	9	**2.27**	0.72	-1.3	0.64	-0.9	0.64	Unfit
B13	32	-1.62	1.02	0.2	0.71	-0.1	0.3	Fit
B14	23	0.15	1	0.1	0.89	-0.2	0.44	Fit
B15	17	1.01	0.93	-0.4	0.87	-0.5	0.52	Fit
B16	25	-0.16	0.81	-1	0.65	-0.9	0.57	Fit
B17	26	-0.32	0.84	-0.8	0.68	-0.8	0.53	Fit
B18	24	0	0.89	-0.6	0.8	-0.5	0.51	Fit
B19	24	0	1.19	1.1	1.26	0.8	**0.27**	Unfit
B20	25	-0.16	0.88	-0.6	1.08	0.3	0.46	Fit
B21	19	0.73	0.89	-0.7	0.81	-0.8	0.54	Fit
B22	26	-0.32	0.79	-1.1	0.61	-1	0.57	Fit

Item	Total score	b_i	Infit		Outfit		PMC	Category
			MNSQ	ZSTD	MNSQ	ZSTD		
B23	32	-1.62	1.28	0.8	**1.94**	1.2	**-0.02**	Unfit
B24	26	-0.32	1	0.1	1.02	0.2	0.38	Fit
B25	22	0.3	0.89	-0.7	0.81	-0.6	0.53	Fit
B26	22	0.3	0.93	-0.4	0.8	-0.6	0.51	Fit
B27	12	1.75	1.11	0.6	1.22	0.8	0.37	Fit
B28	21	0.45	0.99	0	0.89	-0.3	0.47	Fit
B29	22	0.3	1.17	1.1	1.17	0.7	0.31	Fit
B30	27	-0.49	0.87	-0.6	0.67	-0.7	0.5	Fit
Mean	23.5	0.00	0.99	0.00	1.09	0.1		
S.D.	5.5	0.92	0.17	0.9	0.58	1.1		

Table 2 informed that in the Indonesian PAS instrument for class V at MIN 1 East Belitung for the odd semester of the 2023/2024 academic year, there were 77% of the questions that were fit and the remaining 23% of the questions were not fit. The difficulty level index for Indonesian PAS questions ranges from -1.62 to +2.27 with an average of 0 and a standard deviation (S.D.) of 0.92. These results indicated that the Indonesian PAS questions had a moderate level of difficulty. Items that met the criteria for good/fit items were item numbers 1, 2, 4, 5, 8, 9, 19, 13, 14, 15, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, and 30 while those that did not meet the criteria were items number 3, 6, 7, 11, 12, 19, and 23.

In addition, in items number 6 and 7, it appeared that both items had the same level of difficulty, namely -0.49 with the number of students who answered correctly as many as 27 people. These two items met the criteria for a good level of difficulty because they had a difficulty index in the range -2 <bi<+2. However, the two questions had a very small contribution to measuring Indonesian students' skills. This could be seen from Pt. The *measure correlation* between the two was very low, namely 0.16 and 0.18. Therefore, questions number 6 and 7 were presented in figure 4.



Teks berikut untuk soal nomor 6 dan 7

Di suatu pagi yang cerah, masyarakat Desa Cendana sedang sibuk mempersiapkan festival tahunan. **Persiapan** festival telah mereka lakukan dari beberapa bukan sebelumnya. Mereka bekerja keras untuk mendekorasi jalan-jalan desa dengan warna-warni yang ceria. Di tengah keriuhan itu, seorang pemuda bernama Agus dengan penuh semangat memimpin kelompoknya. Ia memperhatikan setiap detail dekorasi (.....) memberikan petunjuk kepada rekan-rekannya. Agus bertekad untuk membuat festival kali ini menjadi yang terbaik.

6. Kata dasar dari kata yang dicetak tebal adalah

A. Bersiap	B. Siap
C. Persiap	D. Siapan

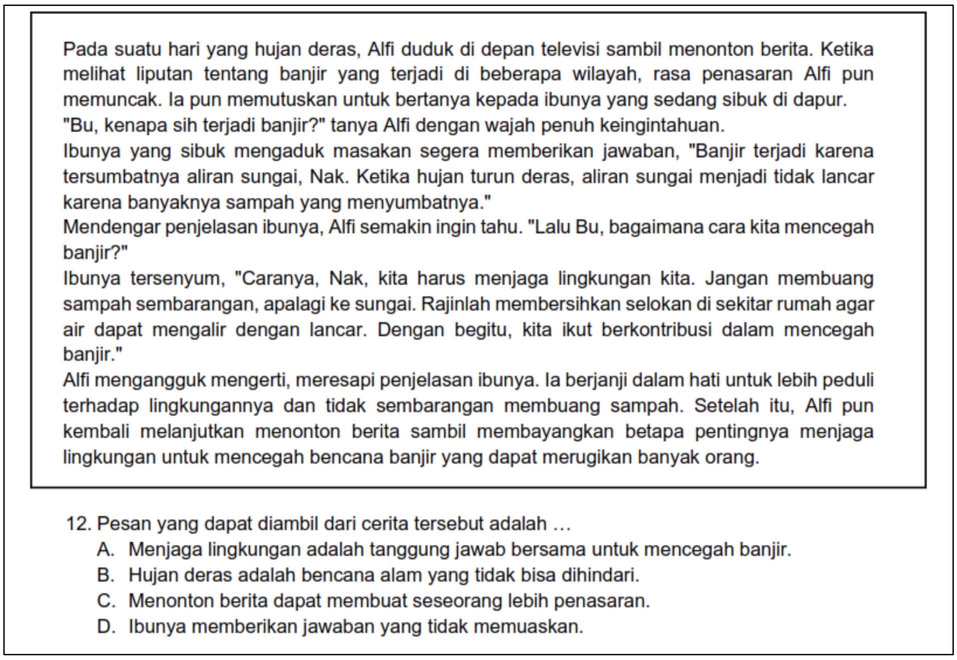
7. Kata hubung yang tepat untuk melengkapi teks tersebut adalah ...

A. tetapi	B. atau
C. maka	D. dan

Figure 4. Indonesian PAS questions number 6 and 7

Based on Figure 4, questions number 6 and 7 came from the same stimulus, namely in the form of text on the topic of annual festivals. The case that occurred in questions number 6 and 7 is understandable because both were prepared from the same main questions so it was natural that they had characteristics that were not much different. These two questions needed to be improved, especially in the subject matter or question stimulus by adjusting the indicators being measured.

In the case of question number 12, it was different from the case of question number 6 and 7. Question number 12 met the criteria for Outfit Mean Square (MNSQ) which was in the range 0.5<MNSQ<1.5, Outfit Z-standard (ZTSD) was in the range -2 <ZSTD<+2, and Point measure correlation in the range 0.3<PMC<0.85 but fails to meet the difficulty level criteria because it had a difficulty index of 2.27. An item difficulty index above +2 indicated that the item had a very high level of difficulty, meaning that the item was too difficult for students to complete (Effatpanah, Baghaei, & Karimi, 2024; Qiu, Peabody, & Bradley, 2024). This can be seen in the total score of 9, meaning that of the 36 students who took Indonesian PAS, only 9 students (25%) were able to answer item number 12 correctly, and the remaining 75% answered incorrectly. Therefore, item number 12 needed to be revised so that it was suitable for use to measure students' abilities. Therefore, question number 12 is shown in figure 12.



Pada suatu hari yang hujan deras, Alfi duduk di depan televisi sambil menonton berita. Ketika melihat liputan tentang banjir yang terjadi di beberapa wilayah, rasa penasaran Alfi pun memuncak. Ia pun memutuskan untuk bertanya kepada ibunya yang sedang sibuk di dapur.

"Bu, kenapa sih terjadi banjir?" tanya Alfi dengan wajah penuh keingintahuan.

Ibunya yang sibuk mengaduk masakan segera memberikan jawaban, "Banjir terjadi karena tersumbatnya aliran sungai, Nak. Ketika hujan turun deras, aliran sungai menjadi tidak lancar karena banyaknya sampah yang menyumbatnya."

Mendengar penjelasan ibunya, Alfi semakin ingin tahu. "Lalu Bu, bagaimana cara kita mencegah banjir?"

Ibunya tersenyum, "Caranya, Nak, kita harus menjaga lingkungan kita. Jangan membuang sampah sembarangan, apalagi ke sungai. Rajinlah membersihkan selokan di sekitar rumah agar air dapat mengalir dengan lancar. Dengan begitu, kita ikut berkontribusi dalam mencegah banjir."

Alfi mengangguk mengerti, meresapi penjelasan ibunya. Ia berjanji dalam hati untuk lebih peduli terhadap lingkungannya dan tidak sembarangan membuang sampah. Setelah itu, Alfi pun kembali melanjutkan menonton berita sambil membayangkan betapa pentingnya menjaga lingkungan untuk mencegah bencana banjir yang dapat merugikan banyak orang.

12. Pesan yang dapat diambil dari cerita tersebut adalah …
    A. Menjaga lingkungan adalah tanggung jawab bersama untuk mencegah banjir.
    B. Hujan deras adalah bencana alam yang tidak bisa dihindari.
    C. Menonton berita dapat membuat seseorang lebih penasaran.
    D. Ibunya memberikan jawaban yang tidak memuaskan.

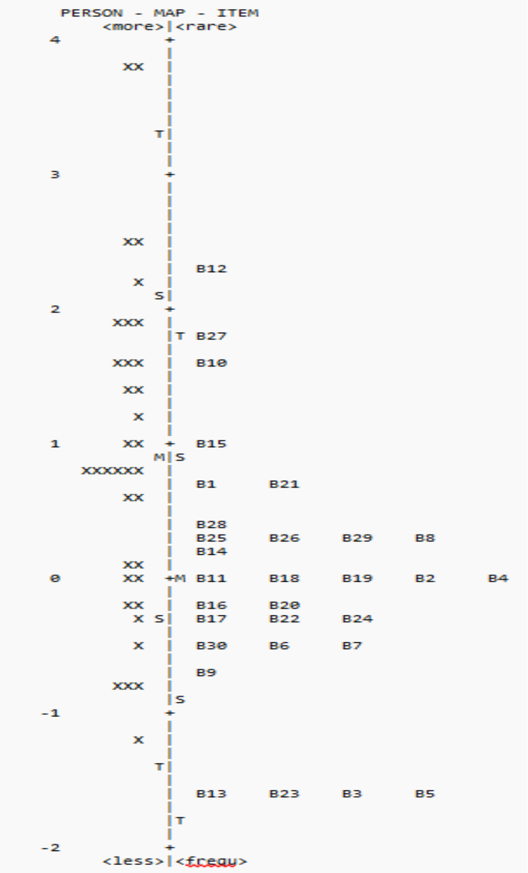
Figure 5. Indonesian PAS question item number 12

Based on the Rasch model approach as explained previously, it could be concluded that valid question items were question items number 1, 2, 4, 5, 8, 9, 19, 13, 14, 15, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, and 30 while items number 3, 6, 7, 11, 12, 19, and 23 were invalid. The results of the reliability estimation showed that the Indonesian PAS questions at MIN 1 Bangka Belitung had good instrument reliability as shown by the item reliability coefficient of 0.77, person reliability of 0.80, and Cronbach's alpha of 0.84.

The item reliability value of 0.77 indicated that the items in the Indonesian PAS instrument had fairly good consistency in measuring individual abilities. This meant that the items could be relied upon to measure the desired abilities (Vrotsou et al., 2023; Zhao, Huen, & Chan, 2017). The person reliability value of 0.80 indicated that the abilities or traits measured were quite consistent among the individuals tested. This value indicated that the instrument was quite reliable in measuring variations in individual abilities (Abdullaev, Shukhratovna, Rasulovna, Umirzakovich, & Staroverova, 2024; Mahtari, Misbah, & Suryati, 2019; Tran et al., 2018). The Cronbach's Alpha value of 0.84 indicates that the instrument has a high level of reliability. Reliability values above 0.70 indicate good reliability (Cronbach, 1951; Tavakol & Dennick, 2011). By fulfilling the validity and reliability parameters based on the Rasch model approach, it could be concluded that the Indonesian PAS instrument for class V at MIN 1 East Belitung was suitable for use to measure Indonesian students' skills, provided that items that were declared invalid were revised or eliminated.

Wright map or person-item map

Wright map, or person-item map was a visual representation that showed the distribution of individual abilities and the level of difficulty of items on the same scale. The Wright map of the results of testing the Indonesian PAS instrument was shown in figure 6.



```
     PERSON - MAP - ITEM
         <more>|<rare>
    4          +
               |
            XX |
               |
               |
               |
               |
              T|
               |
               |
    3          +
               |
               |
               |
               |
            XX |
               |
               |  B12
             X |
              S|
    2          +
           XXX |
               |T B27
               |
           XXX |  B10
               |
            XX |
               |
             X |
               |
    1       XX +  B15
              M|S
        XXXXXX |
               |  B1     B21
            XX |
               |
               |  B28
               |  B25    B26    B29    B8
               |  B14
            XX |
    0       XX +M B11    B18    B19    B2     B4
               |
            XX |  B16    B20
            X S|  B17    B22    B24
               |
             X |  B30    B6     B7
               |
               |  B9
           XXX |
               |S
   -1          +
               |
             X |
               |
              T|
               |
               |  B13    B23    B3     B5
               |
               |T
               |
   -2          +
         <less>|<frequ>
```

Figure 6. Wright map or distribution of student abilities and level of item difficulty

Based on Figure 6, the vertical line in the center represents the logits scale, which was a log-odds unit scale used in the Rasch model to measure both individual ability and item difficulty. This scale ranged from -2 to 4, indicating the variation in individual ability and item difficulty. The left side of the map shown the distribution of individual abilities. Each "X" represents a group of individuals with similar abilities. The higher the "X" position on the scale, the higher the individual's abilities. Conversely, the lower the "X" position, the lower the individual's abilities (Briggs, 2019; Gana & Guillaume Broc, 2019; Petrillo, Cano, McLeod, & Coon, 2015). The right side of the map shown the distribution of item difficulty (B1, B2, B3, etc.) on the same scale. Items located higher on the scale were more difficult, while items located lower were easier (Briggs, 2019).

Figure 6 also shown that there were several students with high abilities located around the logit value of 4 (XX) beyond items that had a high level of difficulty such as items number 12 and 27 which were located at the top of the scale, around the logit value of 3. Medium-ability students were more than high-ability students. Many individuals had moderate abilities which were distributed around the logit value of 0 to 2 as shown by the distribution of scale, around the logit values -1 to -2 there were 4 students as shown by the X distribution. Hence, it was also visible that items that were relatively easy, such as items number 23 and 3, were located at the bottom of the logit scale.

CONCLUSION

The Based on the presentation of the results and discussion, it can be concluded that the PAS question for Indonesian subject in grade V at MIN 1 East Belitung for the 2023/2024 academic year meets the criteria for a good test instrument based on the Rasch model approach. This conclusion is supported by the fact that 77% of the items meet the fit criteria in the Rasch model, while the remaining 23% do not. The reliability of the instrument is demonstrated by an item reliability coefficient of 0.77, indicating good consistency among the test items; a person reliability coefficient of 0.80, reflecting reliable measurement of students' abilities; and a Cronbach’s alpha of 0.84, showing strong internal consistency of the test items. The items that are suitable include numbers 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 20, 21, 22, 24, 25, 26, 27, 28, 29, and 30, while the items that are not fit include numbers 3, 6, 7, 11, 12, 19, and 23. Items that do not fit are unsuitable for assessing students' abilities and need to be revised or removed. Furthermore, using items that do not meet the criteria could result in invalid information. This issue should be a concern for school principals, teachers, and education stakeholders. Improving teacher competence in developing test items is also necessary through seminars, training, or workshops. The contributions of this research include providing a framework for enhancing the validity and reliability of assessment instruments and offering insights into effective assessment practices. This study is limited to multiple-choice questions; therefore, further research should investigate items in descriptive formats or mixed multiple-choice and descriptive formats using the Rasch model approach.

REFERENCES

Abdellatif, H. (2023). Test results with and without blueprinting: Psychometric analysis using the Rasch model. *Educación Médica*, *24*(3), 100802. https://doi.org/10.1016/j.edumed.2023.100802

Abdullaev, D., Shukhratovna, D. L., Rasulovna, J. O., Umirzakovich, J. U., & Staroverova, O. V. (2024). Examining Local Item Dependence in a Cloze Test with the Rasch Model. *International Journal of Language Testing*, *14*(1), 75–81.

Ahsani, E. L. F., & Taqiyah, B. (2024). Implementasi Model Pembelajaran Project Based Learning pada Mata Pelajaran Bahasa Indonesia. *Azkiya*, *9*(1), 130–141. https://doi.org/10.32505/azkiya.v9i1.8443

Akyol, Z., & Garrison, D. R. (2011). Understanding cognitive presence in an online and blended community of inquiry: Assessing outcomes and processes for deep approaches to learning. *British Journal of Educational Technology*, *42*(2), 233–250. https://doi.org/10.1111/j.1467-8535.2009.01029.x

Anderson, L. W., & Krathwohl, D. R. (2001). *A Taxonomy for Learning Teaching and Assessing: A Revision of Bloom's Taxonomy of Education Objectives*. New York: Addison Wesley Longman, Inc.

Arlinwibowo, J., Retnawati, H., & Kartowagiran, B. (2021). Item Response Theory Utilization for Developing the Student Collaboration Ability Assessment Scale in STEM Classes. *Ingenierie Des Systemes d'Information*, *26*(4), 409–415. https://doi.org/10.18280/ISI.260409

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

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

Briggs, D. C. (2019). Interpreting and visualizing the unit of measurement in the Rasch Model. *Measurement*, *146*, 961–971. https://doi.org/10.1016/j.measurement.2019.07.035

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, (16), 297–334.

Danni, R., & Tauratiya, T. (2020). Analisis Kemampuan Berpikir Kritis Mahasiswa Program Studi Hukum Keluarga Islam IAIN Syaikh Abdurrahman Siddik Bangka Belitung. *Tarbawy : Jurnal Pendidikan Islam*, *7*(1), 17–22. https://doi.org/10.32923/tarbawy.v7i1.1191

Danni, R., Wahyuni, A., & Tauratiya, T. (2021). Item Response Theory Approach: Kalibrasi Butir Soal Penilaian Akhir Semester Mata Pelajaran Bahasa Arab. *Arabi : Journal of Arabic Studies*, *6*(1), 93–104. https://doi.org/10.24865/AJAS.V6I1.320

de la Torre, J., & Minchen, N. (2014). Cognitively Diagnostic Assessments and the Cognitive Diagnosis Model Framework. *Psicología Educativa*, *20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

DeMars, C. E. (2018). Classical Test Theory and Item Response Theory. In *The Wiley Handbook of Psychometric Testing* (pp. 49–73). Chichester, UK: John Wiley & Sons, Ltd. Retrieved from http://doi.wiley.com/10.1002/9781118489772.ch2

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Effatpanah, F., Baghaei, P., & Karimi, M. N. (2024). A mixed Rasch model analysis of multiple profiles in L2 writing. *Assessing Writing*, *59*, 100803. https://doi.org/10.1016/j.asw.2023.100803

Fajrianthi, F., Hendriani, W., & Septarini, B. G. (2016). Pengembangan Tes Berpikir Kritis Dengan Pendekatan Item Response Theory. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 45–55. https://doi.org/10.21831/pep.v20i1.6304

Fauziana, A., & Wulansari, A. D. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch. *Jurnal Ibriez : Jurnal Kependidikan Dasar Islam Berbasis Sains*, *6*(1), 10–19. https://doi.org/10.21154/ibriez.v6i1.112

Gana, K., & Guillaume Broc. (2019). *Structural Equation Modeling with Lavaan*. United States: John Wiley & Sons, Ltd.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

Istiqlal, M., Istiyono, E., Widihastuti, W., Sari, D. K., Danni, R., & Safitri, I. (2024). Construction of Mathematics Cognitive Test Instrument of Computational Thinking Model for Madrasah Aliyah Students. *Nazhruna: Jurnal Pendidikan Islam*, *7*(2), 475–492. https://doi.org/10.31538/nzh.v7i2.4425

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (Pysthots) Peserta Didik SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *18*(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Knorn, S., Topalovic, D., & Varagnolo, D. (2022). Redesigning a classic control course using constructive alignment, student-centred teaching, and continuous assessment. *IFAC-PapersOnLine*, *55*(17), 180–185. https://doi.org/10.1016/j.ifacol.2022.09.312

Mahtari, S., Misbah, M., & Suryati, S. (2019). Analysis of the Ability of High School Students to Solving Science Literacy Questions Based on the Rasch Model. *Kasuari: Physics Education Journal (KPEJ)*, *2*(1), 11–16. https://doi.org/10.37891/KPEJ.V2I1.61

Mardapi, D. (2016). *Pengukuran penilaian dan evaluasi pendidikan* (2nd ed.). Yogyakarta: Nuha Litera.

Mustafa, N., Khairani, A. Z., & Ishak, N. A. (2021). Calibration of the science process skills among Malaysian elementary students: A Rasch model analysis. *International Journal of Evaluation and Research in Education (IJERE)*, *10*(4), 1344–1351. https://doi.org/10.11591/ijere.v10i4.21430

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and Rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Qiu, C., Peabody, M. R., & Bradley, K. D. (2024). Exploring Construct Measures Using Rasch Models and Discretization Methods to Analyze Existing Continuous Data. *Measurement: Interdisciplinary Research and Perspectives*, *22*(1), 108–120. https://doi.org/10.1080/15366367.2023.2210358

Ramadhan, S., Mardapi, D., Prasetyo, Z. K., & Utomo, H. B. (2019). The development of an instrument to measure the higher-order thinking skill in Physics. *European Journal of Educational Research*, *8*(3), 743–751.

Ramadhan, S., Sumiharsono, R., Mardapi, D., & Prasetyo, Z. K. (2020). The Quality of Test Instruments Constructed by Teachers in Bima Regency, Indonesia: Document Analysis. *International Journal of Instruction*, *13*(2), 507–507. https://doi.org/10.29333/iji.2020.13235a

Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya*. Yogyakarta: Nuha Medika.

Retnawati, H. (2016). *Validitas, reliabilitas dan karakteristik butir*. Yogyakarta: Nuha Medika.

Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and Assessment in Education* (2nd ed.). Upper Saddle River, NJ: Pearson.

Rezi, L. (2024, March 23). *Assessment system in MIN 1 East Belitung*.

Sarea, M. S., & Ruslan, R. (2019). Karakteristik Butir Soal: Classical Test Theory Vs Item Response Theory? *DIDAKTIKA : Jurnal Kependidikan*, *13*(1), 1–16. https://doi.org/10.30863/didaktika.v13i1.296

Stiggins, R., & Chappuis, J. (2012). *Introduction to student invoved assessment for learning* (6th ed.). Boston: Addison Wesley.

Sudjana, N. (2016). *Penilaian Hasil Belajar Proses Belajar Mengajar*. Bandung: Remaja Rosdakarya.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Bandung: Trim Komunikata.

Susetyo, B. (2015). *Prosedur Penyusunan dan Analasis Tes*. Bandung: PT Refika Aditama.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tran, V. D., Dorofeeva, V. V., & Loskutova, E. E. (2018). Development and validation of a scale to measure the quality of patient medication counseling using Rasch model. *Pharmacy Practice*, *16*(4), 1327. https://doi.org/10.18549/PharmPract.2018.04.1327

Vrotsou, K., Subiza-Pérez, M., Lertxundi, A., Vergara, I., Marti-Carrera, I., Ochoa de Retana, L., … Ibarluzea, J. (2023). Environmental health knowledge of healthcare professionals: Instrument development and validation using the Rasch model. *Environmental Research*, *235*, 116582. https://doi.org/10.1016/j.envres.2023.116582

Wilson, M. (2023). *Constructing Measures An Item Response Modeling Approach* (2nd ed.). New York: Routledge.

Xu, T., & Stone, C. A. (2012). Using IRT Trait Estimates Versus Summated Scores in Predicting Outcomes. *Educational and Psychological Measurement*, *72*(3), 453–468. https://doi.org/10.1177/0013164411419846

Zeigenfuse, M. D., Batchelder, W. H., & Steyvers, M. (2020). An item response theory model of matching test performance. *Journal of Mathematical Psychology*, *95*, 102327–102327. https://doi.org/10.1016/j.jmp.2020.102327

Zhao, Y., Huen, J. M. Y., & Chan, Y. W. (2017). Measuring Longitudinal Gains in Student Learning: A Comparison of Rasch Scoring and Summative Scoring Approaches. *Research in Higher Education*, *58*(6), 605–616. https://doi.org/10.1007/s11162-016-9441-z