AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 14, No. 1, 2025, 299 – 310

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

OPTIMIZING NUMERACY ASSESSMENTS: A DETAILED ITEM ANALYSIS FOR PHASE D STUDENTS

Sindi Amelia[1*], Indah Widiati[2], Gusri Yadrika[3]

[1,2] Universitas Islam Riau, Pekanbaru, Indonesia
[3] Universitas Riau, Pekanbaru, Indonesia
**Corresponding author. Jl. Kaharuddin Nasution No. 113, 28284, Pekanbaru, Indonesia.*
E-mail: *sindiamelia88@edu.uir.ac.id* [1*)]
indahwidiati@edu.uir.ac.id [2)]
gusri.yadrika6518@grad.unri.ac.id [3)]

Received 11 June 2024; Received in revised form 17 March 2025; Accepted 30 April 2025

Abstract

Compared to literacy skills, the numeracy skills of Indonesian students are more alarming. Therefore, best practices that support the enhancement of students' numeracy skills are urgently needed as an effort to improve the academic performance of Indonesian students. One effective practice is giving numeracy problems routinely to train students in dealing with numeracy-based problems.. Developing high-quality numeracy questions requires a systematic and scientific approach. In previous research, a numeracy instrument for Phase D students was developed and validated through a series of qualitative activities (Self-Evaluation, Expert Review, One-to-One, and Small Group). To further improve the quality of the numeracy questions for Phase D students, this study will continue with quantitative activities, specifically through the Field Test phase. Aim of this study is to analyse the numeracy questions for Phase D students using data analysis techniques with the aid of the Quest program. This study resulted in 21 valid numeracy questions with an ideal difficulty distribution (19% difficult, 62% moderate, and 19% easy), and a high reliability score (0.92). These questions can be used at the middle school level or Phase D as diagnostic, formative, or summative assessments to measure students' numeracy skills.

Keywords: numeracy skill; quest program; student phase d.

Abstrak

Dibandingkan dengan kemampuan literasi, kemampuan numerasi peserta didik Indonesia lebih memprihatinkan. Oleh karena itu, praktik yang dapat mendukung peningkatan kemampuan numerasi sangat dibutuhkan sebagai bentuk usaha untuk meningkatkan prestasi akademik peserta didik. Salah satunya adalah pemberian soal numerasi secara rutin untuk melatih peserta didik dalam menghadapi soal-soal berbasis kemampuan numerasi. Penyusunan soal numerasi yang berkualitas perlu melalui tahapan pengembangan yang ilmiah. Pada penelitian sebelumnya, telah dikembangkan instrumen soal numerasi untuk peserta didik fase D yang telah teruji validitas dan kepraktisannya melalui serangkaian kegiatan kualitatif, yaitu Self-Evaluation, Expert Review, One-to-one, dan Small Group. Untuk menyempurnakan kualitas soal numerasi bagi peserta didik fase D, penelitian dilanjutkan dengan kegiatan kuantitatif, yakni melalui tahapan Field Test. Tujuan dari penelitian ini adalah untuk menganalisis butir soal numerasi untuk peserta didik fase D dengan teknik analisis data menggunakan bantuan program Quest. Penelitian ini menghasilkan 21 butir soal numerasi yang valid dengan tingkat kesukaran yang ideal, yakni 19% soal sulit, 62% soal sedang, dan 19% soal mudah, serta tingkat reliabilitas yang baik sekali (0,92). Butir soal ini dapat digunakan di jenjang SMP atau Fase D sebagai asesmen diagnostik, formatif, maupun sumatif untuk mengukur tingkat kemampuan numerasi peserta didik.

***Kata kunci**: kemampuan numerasi; peserta didik fase d; program quest.*

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

INTRODUCTION

Numeracy is simply viewed dichotomously as testing whether a person can perform basic arithmetic or not (Parnis & Petocz, 2016). Mastering numeracy skills means having the ability to think critically in processing data, making decisions, and solving problems effectively (Yustitia, Kusmaharti, & Wardani, 2025).

International assessments such as PISA (Programme for International Student Assessment) have consistently shown that Indonesian students struggle with basic numeracy skills. In line with this, the national assessment, namely Minimum Competency Assessment (AKM), also shows that students' numeracy is still relatively low (Rosnelli & Ristiana, 2023).

One crucial effort to enhance students' numeracy skills is through regular practice with numeracy questions. An activity that can increase numeracy scores is intensive training in answering numeracy-related questions (Ismawati, Hersulastuti, Amertawengrum, & Anindita, 2023; Kholid et al., 2022). As a tool for practice, teachers need a collection of well-developed numeracy questions. Therefore, the development of high-quality numeracy questions is essential.

The various stages required to produce a high-quality question instrument include expert validation, readability testing (both limited and in small groups), and item analysis. Item analysis aims to assess the validity, reliability, discriminative power, and difficulty level of the questions. In previous research, a set of questions that were valid (based on expert judgment) and practical (through two stages of readability testing with students) was developed. To further test the quality, the instrument must undergo a final stage of quantitative analysis.

Referring to the Minimum Competency Assessment (AKM), which serves as a benchmark for measuring the quality of each school in Indonesia, several forms of numeracy questions are provided: essay, short answer, multiple choice, complex multiple choice, and matching. These forms can be categorized into dichotomous and polytomous questions.

The Quest program, an application used for calculating and analyzing question items, has the advantage of being able to analyse both dichotomous and polytomous questions. Additionally, this program can estimate both item groups and respondent groups, making it the primary choice for researchers to produce truly high-quality numeracy questions. The objective of this study is to determine the quality of numeracy question items (including essay, short answer, multiple choice, complex multiple choice, and matching) for Phase D students through quantitative analysis using the Quest program.

Phase D students are those in grades 7, 8, and 9 in Junior High School. One of the competencies measured in the basic education competency standards (SKL) is the ability to demonstrate numeracy skills by reasoning using mathematical concepts, procedures, facts, and tools to solve problems related to oneself, the immediate environment, and the wider community (Kemendikbudristek RI, 2022).

In relation to numeracy, Phase D students can answer numeracy questions based on domain components aligned with the Learning Outcomes (CP) set by the government through the Kurikulum Merdeka.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

Table 1 summarizes several studies related to the analysis of numeracy question items.

Table 1. State-of-the-Art Analysis of Numeracy Question Items

No	Research Title / Author Name (Year)	Research Design	Result
1	Validating an Instrument to Evaluate the Teaching of Mathematics Through Processes / (Alsina, Maurandi, Ferre, & Coronata, 2021)	Stuctural Equation Model	Thirty-five questionnaire items administered to 95 Spanish early years and primary education teachers show a high coefficient and a significant p-value.
2	Validation of a Digital Tool for Diagnosing Mathematical Proficiency / (Junpeng et al., 2020)	Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM)	The instrument is validated based on three arguments: validity, reliability, and item fit, making it suitable for use as a formative test in schools.
3	Estimation of college students' ability on Real Analysis course using Rasch model / (Isnani, Utami, Susongko, & Lestiani, 2019)	Rasch Model with Quest Program	100% of essay questions in Real Analysis final exam is categorized as difficult.
4	Learning number patterns through computational thinking activities / (Chan et al., 2021)	Rasch Model	Eight items consisting of arithmetic sequence, quadratic sequence, and geometric sequence materials have good construct validity adn the items were productive so that they are acceptable for a good measurement.
5	Pengembangan tes kemampuan pemecahan masalah dan penalaran matematika siswa SMP kelas VIII (Sinaga, 2016)	Rasch Model with Quest Program	The developed test instrument for mathematical problem-solving and reasoning skills is, overall, appropriate and valid for use.

Among the five aforementioned studies, item analysis tends to focus on either the elementary school level or the university level. For Phase D, research subjects are only available in grade 8. However, the study's question material covers all levels within Phase D. Furthermore, the item analysis in this study encompasses not only one type of question but includes all question formats present in the Minimum Competency Assessment (AKM). Thus, the purpose of this study is to examine numeracy questions for Phase D

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

students by applying data analysis techniques with the support of the Quest program.

METHODS

In general, this research constitutes a series of item development studies utilizing a formative evaluation design. This study is at the Field Test stage, where in previous research, a numeracy item instrument was obtained, which was both valid (92%) and practical (88%) (Amelia, Widiati, & Yadrika, 2023) through the stages of self-evaluation, expert review, one-to-one, and small group, as illustrated in Figure 1.

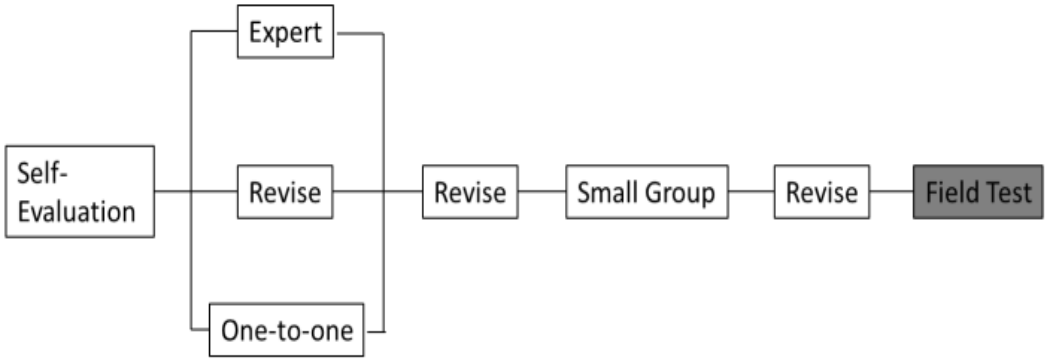


Figure 1. Desain Formative Evaluation

Specifically, this research is evaluative in nature, employing a quantitative descriptive approach. The evaluation focuses on numeracy items for Phase D students, assessing their quality through quantitative item analysis. The Quest program is used as a tool for this analysis. Items are considered to be of good quality if they meet the established criteria for item evaluation.

In analysing items, computer programs are commonly used to facilitate the calculation process. One such program is Quest. By utilizing the Quest program, users can effectively and quickly analyse items.

The advantage of this program is its capability to analyse both dichotomous and polytomous data. The program's output allows for the analysis of items from various perspectives within classical theory, such as reliability, difficulty level, discrimination, and distractor items. Additionally, it can analyse the difficulty level of the Rasch model (Reffiane, Sudarmin, Wiyanto, & Saptono, 2021).

The numeracy questions tested quantitatively consist of 30 items, and the percentage distribution of questions based on their components can be seen in Table 2. From the Table 2, 30 numeracy items are presented, distributed across each component and subcomponent of numeracy questions. The proportion of the distribution of numeracy questions is aimed to approximate the proportion of AKM numeracy questions (Pusat Asesmen dan Pembelajaran, 2020).

These questions were administered to 32 Phase D students at SMPN 34 Pekanbaru, then assessed based on their answer alternatives for subsequent analysis using the Quest program.

The output of the Quest program includes item validity estimates, difficulty level estimates, item passing

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

estimates, and reliability estimates (Rizbudiani, Jaedun, Rahim, & Nurrahman, 2021) (see Fig. 2).

Table 2. The Proportion of Numeracy Items Before Analysis

Components	Subcomponents	Items	Proportion
Domain	Number	11, 12, 15, 16, 17, 18	20%
	Measurement and Geometry	4, 5, 6, 7, 26, 27	20%
	Data and Uncertainty	19, 20, 21, 24, 25, 28, 29, 30	27%
	Algebra	1, 2, 3, 8, 9, 10, 13, 14, 22, 23	33%
Context	Personal	1, 2, 3, 8, 9, 10, 11, 12, 24, 25	33%
	Socio-Cultural	13, 14, 15, 16, 17, 18	20%
	Scientific	4, 5, 6, 7, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30	47%
Cognitive Level	Understanding	8, 9, 10, 11, 12, 19, 20, 21, 25, 28, 29, 30	40%
	Application	1, 2, 3, 4, 5, 13, 14, 22, 23	30%
	Reasoning	6, 7, 15, 16, 17, 18, 24, 26, 27	30%
Question Format	Essay	4, 5, 6, 7, 12, 26, 27	23%
	Short Answer	1, 2	7%
	Multiple Choice	11, 13, 14, 20, 21, 22	20%
	Complex Multiple Choice	3, 8, 9, 15, 16, 17, 18, 19, 24, 28, 29, 30	40%
	Matching	10, 23, 25	10%

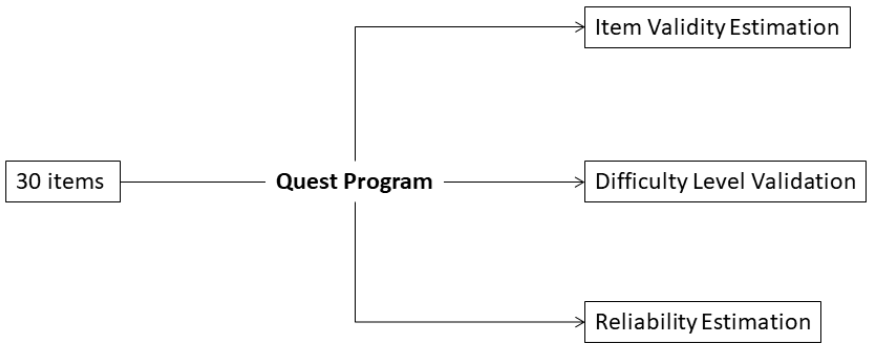


Figure 2. Scheme of the Item Analysis Process Using the Quest Program

In Item Validity Estimation, based on the Rasch Model, the validity of the analysed items can be assessed using the INFIT MNSQ and OUTFIT t output values (Saryanto, Sumiharsono, Ramadhan, & Suprapto, 2020). An item is considered valid if the INFIT MNSQ value falls within the range of 0.5 – 1.5 (Aryadoust, Ng, & Sayama, 2021) and the OUTFIT t value is less than 2.0 (Alhadabi & Aldhafri, 2021; Rusyid et

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

al., 2024; Sukarelawan, Jumadi, Kuswanto, & Thohir, 2021).

The item estimate (Threshold) analysis can also be used to determine the difficulty level of the item. The difficulty levels are categorized as follows: 1) $b > 2$ (very difficult); 2) $1 < b \leq 2$ (difficult); 3) $-1 < b \leq 1$ (moderate); 4) $-2 < b \leq -1$ (easy); and 5) $b < -2$ (very easy) (Dewi, Damio, & Sukarno, 2023).

The criteria for Rasch model reliability values are as follows: 1) $<$ 0.67 (weak); 2) $0.67 - 0.80$ (sufficient); 3) $0.81 - 0.90$ (good); 4) $0.91 - 0.94$ (very good); and > 0.94 (excellent) (Sumintono & Widhiarso, 2015).

RESULTS AND DISCUSSION

1. Item Validity Estimation

The validity results of the numeracy items based on the INFIT MNSQ and OUTFIT t values are shown in Table 3.

Table 3. Recapitulation of Numeracy Item Validity

Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation	Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation
1	1,33	1,2	Valid	16	0,73	-1,1	Valid
2	0,85	-0,5	Valid	17	0,76	-0,9	Valid
3	0,67	-0,8	Valid	18	-	-	Not Valid
4	0,79	-0,7	Valid	19	1,12	1,0	Valid
5	0,86	-0,7	Valid	20	1,61	3,2	Not Valid
6	0,85	-0,8	Valid	21	1,42	3,0	Not Valid
7	0,78	-0,4	Valid	22	0,97	-0,6	Valid
8	1,19	0,4	Valid	23	0,48	-2,3	Valid*
9	1,19	0,4	Valid	24	1,05	1,0	Valid
10	1,58	1,1	Valid*	25	-	-	Not Valid
11	0,81	-0,2	Valid	26	-	-	Not Valid
12	0,82	-0,7	Valid	27	-	-	Not Valid
13	-	-	Not Valid	28	1,14	1,2	Valid
14	1,32	3,1	Not Valid	29	0,60	-1,5	Valid
15	0,72	-1,1	Valid	30	-	-	Not Valid

Note: *: valid with consideration

Table 3 provides information about the validity of each numeracy item. The INFIT MNSQ values of the 30 numeracy items above range from $0.48 - 1.58$ and the OUTFIT t values range from -2.3 to 3.2. This means that there are numeracy items that are not valid according to the Rasch model, namely items 10, 14, 20, 21, and 23. In the Rasch model, items that cannot be answered by all respondents are not counted and are automatically considered invalid, such as items 13, 18, 25, 26, 27, and 30.

In this analysis, it was found that some invalid items (10, 23, and 25) are matching type questions, which means that if all of these questions are eliminated, there will be no matching type questions in this numeracy item instrument. However, to familiarize students with practicing numeracy questions in AKM format, all question components must be fulfilled.

The INFIT MNSQ scores for item 10 (1.58) and item 23 (0.48) indicate that they are less productive as measurement instruments, but they do

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

not decrease data quality, although item 23 specifically can affect reliability scores (Adi, Amaruddin, Adi, & A'yun, 2022). The INFIT MNSQ scores for these two items also have a small difference from the validity category threshold (0.5 – 1.5). Meanwhile, in terms of OUTFIT t scores, both of these items fall into the fit category.

Considering these factors, items number 10 and 23 need to be reviewed (Erfan, Maulyda, Ermiana, Hidayati, & Widodo, 2020) or not discarded. However, to improve the quality of these two items, minor revisions are needed, such as improving the wording of the questions to facilitate students' understanding.

2. Difficulty Level Estimation

The analysis results of the difficulty level of numeracy questions for Phase D students can be seen in Table 4.

Table 4. Recapitulation of Numeracy Question Item Difficulty Levels

Item	Threshold Value	Interpretation	Item	Threshold Value	Interpretation
1	-0,74	Moderate	16	-0,65	Moderate
2	-0,20	Moderate	17	-0,84	Moderate
3	0,34	Moderate	18	-	-
4	0,44	Moderate	19	1,93	Difficult
5	0,03	Moderate	20	-0,46	Moderate *
6	2,26	Very Difficult	21	-0,46	Moderate *
7	1,33	Difficult	22	0,63	Moderate
8	-1,34	Easy	23	-0,11	Moderate
9	-1,34	Easy	24	-0,65	Moderate
10	-0,55	Moderate	25	-	-
11	-1,74	Easy	26	-	-
12	-1,07	Easy	27	-	-
13	-	-	28	1,93	Difficult
14	1,56	Difficult*	29	0,34	Moderate
15	-0,65	Moderate	30	-	-

Note: *: not valid

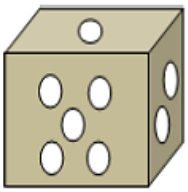
Table 4 presents information on the difficulty levels of numeracy question items. After excluding invalid items (13, 14, 18, 20, 21, 25, 26, 27, and 30), it was found that 4 (19%) items are categorized as very difficult and difficult, 13 (62%) items are moderately difficult, and 4 (19%) items fall into the easy category. This proportion of difficulty levels is considered ideal, with the number of difficult and easy questions together accounting for 19% and the remaining 62% falling into the moderate category.

Item number 6 is an essay question with a maximum score of 2. This item is categorized as very difficult, as only 3 out of 34 respondents scored 1, even though none achieved the perfect score. The numeracy component of question number 6 is an essay format, with reasoning as the cognitive level, falling under the measurement and geometry domain, and using a scientific context, as shown in Fig 3.

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 14, No. 1, 2025, 299 – 310
ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

DICE PAINTING



The picture beside shows a cube-shaped dice with an edge length of 1 cm. Each side of the dice is painted white and features a circle with a diameter of 4 mm.

Question 6: Dice Painting
Find the surface area of the dice that is not white on the 1 and 6-edged sides!

Figure 3. Question Item Number 6 (Very Difficult Category)

This question item involves reasoning, and during the validation stage with experts, the validators recommended increasing its cognitive level compared to the previous question design. Therefore, at this estimation stage, item 6 has not been removed.

3. Reliability Estimation

The reliability of item estimate for multiple-choice questions is 0.92 (very good), and the respondent reliability is 0.81 (good). This means that the respondent reliability is lower than the item reliability. This can occur for several reasons, including respondents answering questions carelessly and the sample size being less than 100 respondents, specifically 32 respondents.

From the results of the three estimations above, 21 valid items were obtained with varying difficulty levels and very good item reliability. The distribution of these 21 numeracy items for Phase D students is presented in the Table 5.

Table 5. Proportion of Numeracy Question Items After Analysing

Components	Subcomponents	Items	Proportion
Domain	Number	11, 12, 15, 16, 17	24%
	Measurement and Geometry	4, 5, 6, 7	19%
	Data dan Uncertainty	19, 24, 28, 29	19%
	Algebra	1, 2, 3, 8, 9, 10, 22, 23	38%
Context	Personal	1, 2, 3, 8, 9, 10, 11, 12, 24	43%
	Socio-Cultural	15, 16, 17	14%
	Scientific	4, 5, 6, 7, 19, 22, 23, 28, 29	43%
Cognitive Level	Understanding	8, 9, 10, 11, 12, 19, 28, 29	38%
	Application	1, 2, 3, 4, 5, 22, 23	33%
	Reasoning	6, 7, 15, 16, 17, 24	29%
Question Format	Essay	4, 5, 6, 7, 12	24%
	Short Answer	1, 2	9,3%
	Multiple Choice	11, 22	9,3%
	Complex Multiple Choice	3, 8, 9, 15, 16, 17, 19, 24, 28, 29	48%
	Matching	10, 23	9,3%

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

From Table 5, it can be seen that the distribution of numeracy questions for Phase D students changed after several items were dropped following analysis with the Quest program.

This research found that nine discarded consisted of Essay (28%), Multiple Choice (66,7%), Complex Multiple Choice (16,7%), and Matching (33,3%) question types. Multiple-choice questions were the most frequently rejected due to their lack of validity.

In this study, questions that students were unable to answer could not be classified as difficult, moderate, or easy. Additionally, item validity, whether based on INFIT MNSQ or OUTFIT t, was not interconnected with difficulty levels. This finding aligns with previous research (Kan, Bulut, & Cormier, 2019; Van Vo & Csapó, 2021). In other words, invalid questions cannot be categorized as difficult, moderate, or easy.

The use of the Quest Program is relatively simple, as it provides readily available command templates. Users only need to input data into the lightweight application. However, Quest has a limitation in that its reliability calculations apply only to multiple-choice questions. This opens opportunities for future researchers to combine Quest with other formulas to obtain reliability values for all question types.

If the development of numeracy questions has the benefit of improving their quality in terms of fulfilling curriculum standards and assessing the validity and practicality of the questions, then the analysis of numeracy items is a follow-up activity to evaluate the potential effects through the validity and reliability of the questions. Item analysis is a simple and easy to understand (Kim, Cohen, & Eom, 2021), yet it can show which items are useful (Asempapa & Lee, 2025). It also strengthens scientific decisions based on the quantitative analysis of the questions' level of difficulty. Thus, integrating both qualitative and quantitative approaches, educators can create a collection of numeracy questions that are truly valid, practical, and effective for classroom use. These process of item quality control is foundational to research in mathematics education (Ing, Kosko, Jong, & Shih, 2024; Quaigrain & Arhin, 2017).

This numeracy test instrument can be further developed to assess the numeracy skills of students in Phase D and analyse their difficulties in solving numeracy questions.

CONCLUSIONS

The Quest program, which is part of the Rasch model, helps in analyzing the validity and reliability of numeracy questions. Additionally, the Quest program provides an overview of the difficulty level of the questions, making it easier to make scientific decisions to produce good numeracy questions for Phase D learners.

After being analysed using the Quest program, this study produced 21 valid numeracy items with an ideal level of difficulty (19% difficult questions, 62% medium questions, and 19% easy questions), as well as a very good level of reliability (0.92). These 21 items not only fulfill curriculum standards but can also be used accurately at the junior high school level or Phase D as diagnostic, formative, and summative assessments to measure students' numeracy skills.

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

ACKNOWLEDGEMENTS

Thanks to DRTPM Dikti for improving this article, and DPPM UIR as research sponsor with contract number: 486/KONTRAK/P-PT/DPPM-UIR/06-2023.

REFERENCES

Adi, N. R. M., Amaruddin, H., Adi, H. M. M., & A'yun, I. L. Q. (2022). Validity and Reliability Analysis Using the Rasch Model to Measure the Quality of Mathematics Test Items of Vocational High Schools. *Journal of Research and Educational Research Evaluation*, *11*(2), 103–113. doi: https://doi.org/10.15294/jere.v11i2.58835

Alhadabi, A., & Aldhafri, S. (2021). A Rasch Model Analysis of the Psychometric Properties of the Student-Teacher Relationship Scale among Middle School Students. *European Journal of Educational Research, volume–10–2021*(volume–10–issue–2–april–2021), 957–973. doi: 10.12973/eu-jer.10.2.957

Alsina, A., Maurandi, A., Ferre, E., & Coronata, C. (2021). Validating an Instrument to Evaluate the Teaching of Mathematics Through Processes. *International Journal of Science and Mathematics Education*, *19*(3), 559–577. doi: 10.1007/s10763-020-10064-y

Amelia, S., Widiati, I., & Yadrika, G. (2023). Pengembangan Soal Numerasi untuk Peserta Didik Fase D. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(3), 3048. doi: 10.24127/ajpm.v12i3.7236

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. doi: 10.1177/0265532220927487

Asempapa, R. S., & Lee, D. (2025). Preservice teachers' knowledge of math modeling: Initial scale development and validation. *Discover Education*, *4*(1), 68. doi: 10.1007/s44217-025-00458-x

Chan, S.-W., Looi, C.-K., Ho, W. K., Huang, W., Seow, P., & Wu, L. (2021). Learning number patterns through computational thinking activities: A Rasch model analysis. *Heliyon*, *7*(9), e07922. doi: 10.1016/j.heliyon.2021.e07922

Dewi, H. H., Damio, S. M., & Sukarno, S. (2023). Item analysis of reading comprehension questions for English proficiency test using Rasch model. *REID (Research and Evaluation in Education)*, *9*(1), 24–36. doi: 10.21831/reid.v9i1.53514

Erfan, M., Maulyda, M. A., Ermiana, I., Hidayati, V. R., & Widodo, A. (2020). Validity and reliability of cognitive tests study and development of elementary curriculum using Rasch model. *Psychology, Evaluation, and Technology in Educational Research*, *3*(1), 26–33. doi: 10.33292/petier.v3i1.51

Ing, M., Kosko, K. W., Jong, C., & Shih, J. C. (2024). Validity evidence of the use of quantitative measures of students in elementary mathematics education. *School Science and*

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

Mathematics, *124*(6), 411–423. doi: 10.1111/ssm.12660

Ismawati, E., Hersulastuti, H., Amertawengrum, I. P., & Anindita, K. A. (2023). Portrait of Education in Indonesia: Learning from PISA Results 2015 to Present. *International Journal of Learning, Teaching and Educational Research*, *22*(1), 321–340. doi: 10.26803/ijlter.22.1.18

Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *REID (Research and Evaluation in Education)*, *5*(2), 95–102. doi: 10.21831/reid.v5i2.20924

Junpeng, P., Marwiang, M., Chinjunthuk, S., Suwannatrai, P., Chanayota, K., Pongboriboon, K., … Wilson, M. (2020). Validation of a digital tool for diagnosing mathematical proficiency. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(3), 665. doi: 10.11591/ijere.v9i3.20503

Kan, A., Bulut, O., & Cormier, D. C. (2019). The Impact of Item Stem Format on the Dimensional Structure of Mathematics Assessments. *Educational Assessment*, *24*(1), 13–32. doi: 10.1080/10627197.2018.1545569

Kemendikbudristek RI. (2022). *Peraturan Menteri Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia Nomor 5 Tahun 2022 tentang Standar Kompetensi Lulusan pada Pendidikan Anak Usia Dini, Jenjang Pendidikan Dasar, dan Jenjang Pendidikan Menengah.*

Kholid, M. N., Rofi'ah, F., Ishartono, N., Waluyo, M., Maharani, S., Swastika, A., … Sari, C. K. (2022). What Are Students' Difficulties in Implementing Mathematical Literacy Skills for Solving PISA-Like Problem? *Journal of Higher Education Theory and Practice*, *22*(2), 181–200.

Kim, S.-H., Cohen, A. S., & Eom, H. J. (2021). A note on the three methods of item analysis. *Behaviormetrika*, *48*(2), 345–367. doi: 10.1007/s41237-021-00131-1

Parnis, A. J., & Petocz, P. (2016). Secondary school students' attitudes towards numeracy: An Australian investigation based on the National Assessment Program—Literacy and Numeracy (NAPLAN). *The Australian Educational Researcher*, *43*(5), 551–566. doi: 10.1007/s13384-016-0218-3

Pusat Asesmen dan Pembelajaran. (2020). *Desain Pengembangan Soal AKM*. Jakarta: Balitbang Kemdikbud.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013. doi: 10.1080/2331186X.2017.1301013

Reffiane, F., Sudarmin, Wiyanto, & Saptono, S. (2021). The instrument analysis of students' problem-solving ability on hybrid learning model using ETNO-STEM Approach through Quest Program in COVID-19 Pandemic. *Pegem Journal of Education and Instruction*, *11*(4), 1–8. doi: 10.47750/pegegog.11.04.01

ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

DOI: https://doi.org/10.24127/ajpm.v14i1.10495

Rizbudiani, A. D., Jaedun, A., Rahim, A., & Nurrahman, A. (2021). Rasch model item response theory (IRT) to analyze the quality of mathematics final semester exam test on system of linear equations in two variables (SLETV). *Al-Jabar : Jurnal Pendidikan Matematika*, *12*(2), 399–412. doi: 10.24042/ajpm.v12i2.9939

Rosnelli, R., & Ristiana, P. A. (2023). Independent Curriculum Learning Management to Improve Students' Literacy and Numerical Competence in Schools. *International Journal of Education in Mathematics, Science and Technology*, *11*(4), 946–963. doi: 10.46328/ijemst.3513

Rusyid, H. K., Suryadi, D., Herman, T., Adnan, M., Lutfi, A., & Mukhibin, A. (2024). Rasch modeling approach to measure the quality of algebraic thinking questions for junior high school students. *Beta: Jurnal Tadris Matematika*, *17*(1), 55–71. doi: 10.20414/betajtm.v17i1.652

Saryanto, S., Sumiharsono, R., Ramadhan, S., & Suprapto, E. (2020). The Analysis of Instrument Quality to Measure theStudents Higher Order Thinking Skill in Physics Learning. *Turkish Journal of Science Education*, *17*(4), 520–527. doi: 10.36681/tused.2020.42

Sinaga, N. A. (2016). Pengembangan tes kemampuan pemecahan masalah dan penalaran matematika siswa SMP kelas VIII. *PYTHAGORAS: Jurnal Pendidikan Matematika*, *11*(2), 169. doi: 10.21831/pg.v11i2.10642

Sukarelawan, Moh. I., Jumadi, J., Kuswanto, H., & Thohir, M. A. (2021). The Indonesian Version of the Physics Metacognition Inventory: Confirmatory Factor Analysis and Rasch Model. *European Journal of Educational Research, volume–10–2021*(volume–10–issue–4–october–2021), 2133–2144. doi: 10.12973/eu-jer.10.4.2133

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

Van Vo, D., & Csapó, B. (2021). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: Evidence of validity and latent predictors of item difficulty. *International Journal of Science Education*, *43*(13), 2185–2205. doi: 10.1080/09500693.2021.1957515

Yustitia, V., Kusmaharti, D., & Wardani, I. S. (2025). Students' critical thinking in numeracy problem-solving through moderate self-Efficacy: A mixed-methods study. *Multidisciplinary Science Journal*, *7*(8), 2025410. doi: 10.31893/multiscience.2025410