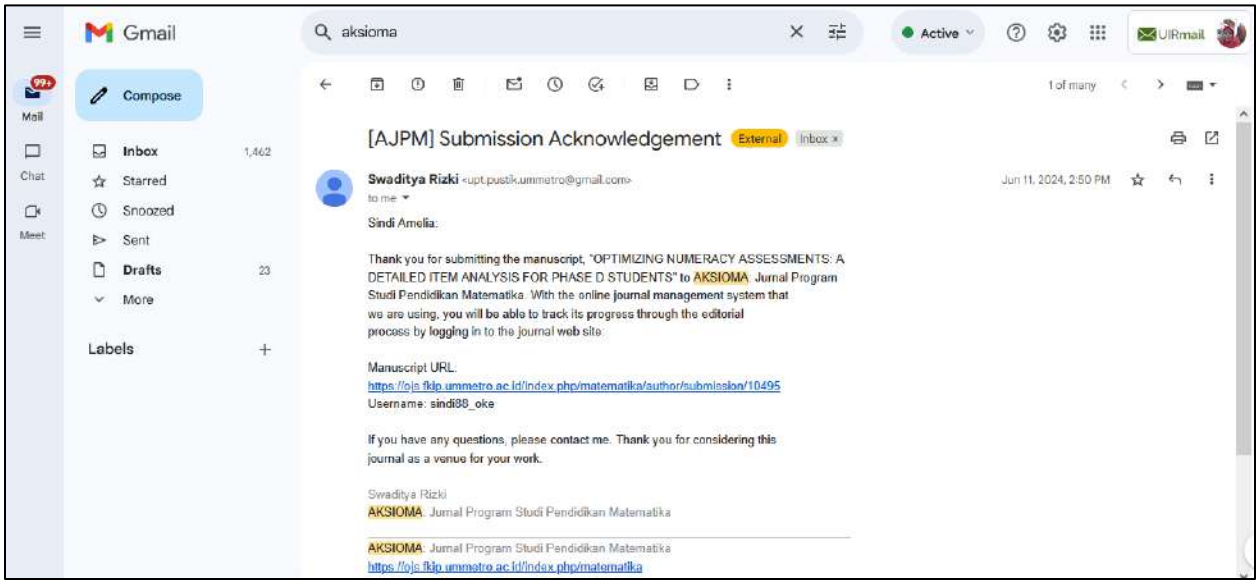
BUKTI KORESPONDENSI
ARTIKEL JURNAL NASIONAL TERAKREDITASI SINTA 2

Judul Artikel	:	Optimizing Numeracy Assessments: A Detailed Item Analysis for Phase D Students
Jurnal	:	AKSIOMA: Jurnal Program Studi Pendidikan Matematika Volume 14, Number 1, 2025, pp. 299-310
Penulis	:	Sindi Amelia, Indah Widiati, Gusri Yadrika

No.	Perihal	Tanggal
1.	Bukti konfirmasi submit artikel dan artikel yang disubmit	11 Juni 2024
2.	Bukti konfirmasi permintaan revisi pertama dan submit revisi yang pertama	19 Februari 2025
3.	Bukti konfirmasi permintaan revisi kedua dan submit revisi yang kedua	19 April 2025
4.	Bukti konfirmasi artikel accepted dan published di web jurnal	30 April 2025

1. Bukti konfirmasi submit artikel dan artikel yang disubmit (11 Juni 2024)



AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 dd Month yy; Received in revised form dd Month yy; Accepted dd Month yy (9pt)

Abstrak

Kemampuan numerasi peserta didik Indonesia yang rendah mestinya menjadi perhatian semua kalangan. Dibandingkan kemampuan literasi, kemampuan numerasi siswa Indonesia lebih memprihatinkan. Untuk itu, praktik-praktik baik yang dapat mendukung peningkatan kemampuan numerasi peserta didik sangat dibutuhkan sebagai bentuk usaha untuk meningkatkan prestasi akademik peserta didik Indonesia. Salah satu praktik baik tersebut adalah dengan rutinnya memberikan soal yang mengukur kemampuan numerasi. Menyusun soal numerasi yang berkualitas, perlu melalui tahapan pengembangan soal yang ilmiah. Pada penelitian sebelumnya, telah dihasilkan instrumen soal numerasi untuk peserta didik fase D yang teruji kevalidan dan kepraktisan melalui serangkaian kegiatan kualitatif (*Self-Evaluation*, *Expert Review*, *One-to-one*, dan *Small Group*). Untuk menyempurnakan kualitas produk soal numerasi untuk peserta didik fase D, maka akan dilanjutkan pada kegiatan kuantitatif, yakni melalui tahapan *Field Test*. Tujuan khusus 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 (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, formative, maupun sumatif untuk mengukur tingkat kemampuan numerasi peserta didik.

Kata kunci: Siswa Fase D; Kemampuan Numerasi; Program Quest.

Abstract

The low numeracy skills of Indonesian students should be a concern for all stakeholders. 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 the regular administration of questions that assess numeracy skills. 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. The specific objective of this study is to analyze 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.*

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

INTRODUCTION

One crucial effort to enhance students' numeracy skills is through regular practice with numeracy questions. 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 analyze 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.

Numeracy Skill

Numeracy is simply viewed dichotomously as testing whether a person can perform basic arithmetic or not (Parnis and Petocz, 2016). Numeracy is not the same as mathematics and is not an alternative to it (Thi et al., 2023). Definitively, numeracy is the ability to develop knowledge and skills by confidently using mathematics to solve practical problems in various aspects of life (Jamil and Khusna, 2021; Directorate of Primary Education, Ministry of Education and Culture, 2021). Similarly, mathematical literacy (numeracy) in PISA is defined as an individual's capacity to formulate, use, and interpret mathematics in diverse contexts (Puspendik, 2019). In other words, numeracy questions tend to refer to PISA-style questions that prioritize reasoning processes over content.

One of the four main policies introduced by the Indonesian Ministry of Education and Culture is to replace the National Examination (UN) with the Minimum Competency Assessment (AKM) and a character survey. The competencies measured in the AKM are literacy and numeracy.

Student Phase D

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

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

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

solve problems related to oneself, the immediate environment, and the wider community (Permendikbudristek, 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.

The following table 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	Development and Validation of Diagnostic Assessment Instrument for Numeracy Skills in 7th Grade / Burgmanis, France, Namsone, & Cakane (2021)	Rasch Model	The diagnostic instrument is suitable for evaluating the numeracy skills of 7th-grade students.
2	Validation of a Digital Tool for Diagnosing Mathematical Proficiency / Junpeng, Marwiang, Chiajunthuk, Suwannatrai, Chanayota, Pongboriboon, Tang, & Wilson (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	Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch / Azzahra, Sunami, & Putranta (2024)	Rasch Model with Quest Program	Seven out of ten multiple-choice items are appropriate, and all items are valid.
4	Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar / Lestari, Hamdu, & Saputra (2023)	Rasch Model	From the multiple-choice questions tested, it was found that 2 questions fall into the difficult category, 8 questions fall into the moderate category, and 1 question falls into the easy category.

From the four aforementioned studies, item analysis tends to focus on the elementary school level. For Phase D, research subjects are only available in grade 7. 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).

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

METHOD

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 the following diagram:

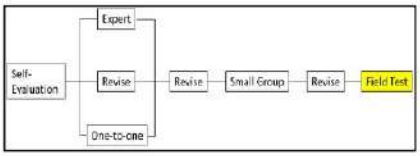


Figure 1. Desain Formative Evaluation

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 analyzing 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 analyze items (Ofianto, 2018).

The advantage of this program is its capability to analyze 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 analyze the difficulty level of the Rasch model (Reffiane et al., 2021).

RESULT AND DISCUSSION

The numeracy questions tested quantitatively consist of 30 items, and the percentage distribution of questions based on their components can be seen in the following table:

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%

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

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

Components	Subcomponents	Items	Proportion
	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%

From the table above, 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 (Assessment and Learning Center, 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 estimates, and reliability estimates (Rizbudiani, Jaedun, Rahim, & Nurrahman, 2021).

1. Item Validity Estimation

In the Rasch model, the validity of the analyzed items can be observed from the output values of INFIT MNSQ and OUTFIT t (Suprapto, Saryanto, Sumiharsono, & Ramadhan, 2020). An item 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 "< 2.0" (Guo, Liu, Hao, Xie, Xiang, Wu, 2020; Bakar et al., 2023; Muslihin, Suryana, Ahman, Suherman, & Dahlan, 2022). The validity results of the numeracy items based on the INFIT MNSQ and OUTFIT t values are shown in the following table:

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

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

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

Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation	Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation
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

The table above 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 not decrease data quality, although item 23 specifically can affect reliability scores (Boone, Yale, Staver, 2014). 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 item estimate (Threshold) can determine the difficulty level of the item. The difficulty level of items can be 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 analysis results of the difficulty level of numeracy questions for Phase D students can be seen in the following table:

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

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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	Mudah	23	-0,11	Moderate
9	-1,34	Mudah	24	-0,65	Moderate
10	-0,55	Moderate	25	-	-
11	-1,74	Mudah	26	-	-
12	-1,07	Mudah	27	-	-
13	-	-	28	1,93	Difficult
14	1,56	Difficult*	29	0,34	Moderate
15	-0,65	Moderate	30	-	-

Note: *: not valid

The table above 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 the image below:

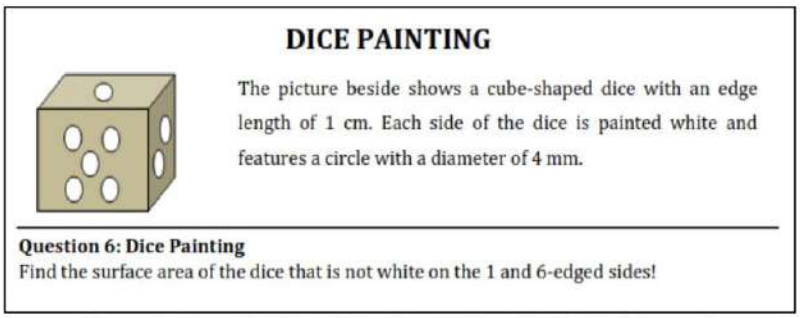


Figure 2. Question Item Number 6 (Very Difficult Category)

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 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). The reliability of item estimate

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

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

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 following table:

Table 6. 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%

From the table above, 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.

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. This also strengthens scientific decisions based on the quantitative analysis of the questions' level of difficulty. Thus, using both approaches (qualitative and quantitative), a collection of numeracy questions that are truly valid, practical, and effective for use in the classroom can be produced.

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

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

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

produce good numeracy questions for Phase D learners.

After being analyzed 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.

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

Azzahra, N. S., Sumarni, S., & Putranta, H. (2024). Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch. QuranicEdu: Journal of Islamic Education, 4(1), 85–94. https://doi.org/10.37252/quranicedu.v4i1.681.

Bakar, S. N. A., Mahmood, N., Ismail, A., Hashim, A., Hamsan, N. Y., Hamid, N. I. N., & Yakzam, R. C. (2023). Validity and Reliability of Competency Analysis Instrument for Cooperative Board Members using Rasch Measurement Model Approach. International Journal of Academic Research in Business and Social Sciences, 13(6), 2408 – 2419. DOI:10.6007/IJARBSS/v13-i6/17543

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). Rasch analysis in the human sciences. In Rasch Analysis in the Human Sciences. https://doi.org/10.1007/978-94-007-6857-4.

G. Burgmanis, I. France, D. Namsone, L. Čakāne (2021) DEVELOPMENT AND VALIDATION OF DIAGNOSTIC ASSESSMENT INSTRUMENT FOR NUMERACY SKILLS IN 7TH GRADE, ICERI2021 Proceedings, pp. 7781-7791.

Chengyao Guo, Yingge Liu, Shengyu Hao, Liang Xie, Guiling Xiang, Yan Wu & Shanqun Li (2020) The Reliability and Validity of the "Activity and Participation" Component in the Brief ICF Core Set for Chronic Obstructive Pulmonary Diseases Based on Rasch Analysis, International Journal of Chronic Obstructive Pulmonary Disease, 15:, 1191-1198, DOI: 10.2147/COPD.S249704

Dewi, H., Damio, S., & 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: https://doi.org/10.21831/reid.v9i1.53514

Direktorat Sekolah Dasar Kemdikbud. (2021). Modul Literasi

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

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

Numerasi di Sekolah Dasar. Jakarta: Kemdikbud.

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: https://doi.org/10.33292/petier.v3i1.51.

Ha Cao Thi, Tuan Anh Le, Bich Tran Ngoc & Thao Phan Thi Phuong (2023) Factors affecting the numeracy skills of students from mountainous ethnic minority regions in Vietnam: Learners' perspectives, Cogent Education, 10:1, DOI: 10.1080/2331186X.2023.2202121.

Jamil, A. F., & Khusna, A. H. (2021). Pengembangan Asesmen Berorientasi Kontekstual untuk Meningkatkan Kemampuan Literasi Matematis dan Numerasi Mahasiswa. Jurnal Ilmiah Mandala Education, 7(4), 78-86. DOI: http://dx.doi.org/10.58258/jime.v7i4.2385.

Junpeng, P., Marwiang, M., Chiajunthuk, S., Suwannatrai, P., Chanayota, K., Pongboriboon, K., Tang, K. N., & Wilson, M. (2020). Validation of a Digital Tool for Diagnosing Mathematical Proficiency. *International Journal of Evaluation and Research in Education (IJERE)*. 9(3). 665-674. DOI: 10.11591/ijere.v9i3.20503

Lestari, T. D., Hamdu. G., & Saputra, E. R., (2023). Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar. *PENDAS: Jurnal Ilmiah Pendidikan Dasar*. 8(2), 2489-2503. DOI: https://doi.org/10.23969/jp.v8i2.9270.

Muslihin, H. Y., Suryana, D., Ahman., Suherman, U., & Dahlan, T. H. (2022). Analysis of the reliability and validity of the self-determination questionnaire using rasch model. International Journal of Instruction, 15(2), 207-222. https://doi.org/10.29333/iji.2022.15212a

Ofianto, O. (2018). Analysis Of Instrument Test Of Historical Thinking Skills In Senior High School History Learning With Quest Programs. Indonesian Journal of History Education, 6(2), 184-192. Retrieved from https://journal.unnes.ac.id/sju/index.php/ijhe/article/view/27648.

Parnis, A.J., Petocz, P. Secondary school students' attitudes towards numeracy: an Australian investigation based on the National Assessment Program—Literacy and Numeracy (NAPLAN). Aust. Educ. Res. 43, 551–566 (2016). https://doi.org/10.1007/s13384-016-0218-3.

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

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

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

Pusat Asesmen dan Pembelajaran. (2020). AKM dan Implikasinya ke Pembelajaran. Jakarta: Balitbang Kemdikbud.

Pusat Penilaian Pendidikan. (2019). Pendidikan di Indonesia: Belajar dari Hasil PISA 2018. Jakarta: Balitbang Kemdikbud.

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

Reffiane F, Sudarmin, Wiyanto, Saptono S (2021). Developing an Instrument to Assess Students' Problem-Solving Ability on Hybrid Learning Model Using Ethno-STEM Approach through Quest Program. Pegem Journal of Education and Instruction, Vol. 11, No. 4, 2021, 1-8. DOI: 10.47750/pegegog.11.04.01.

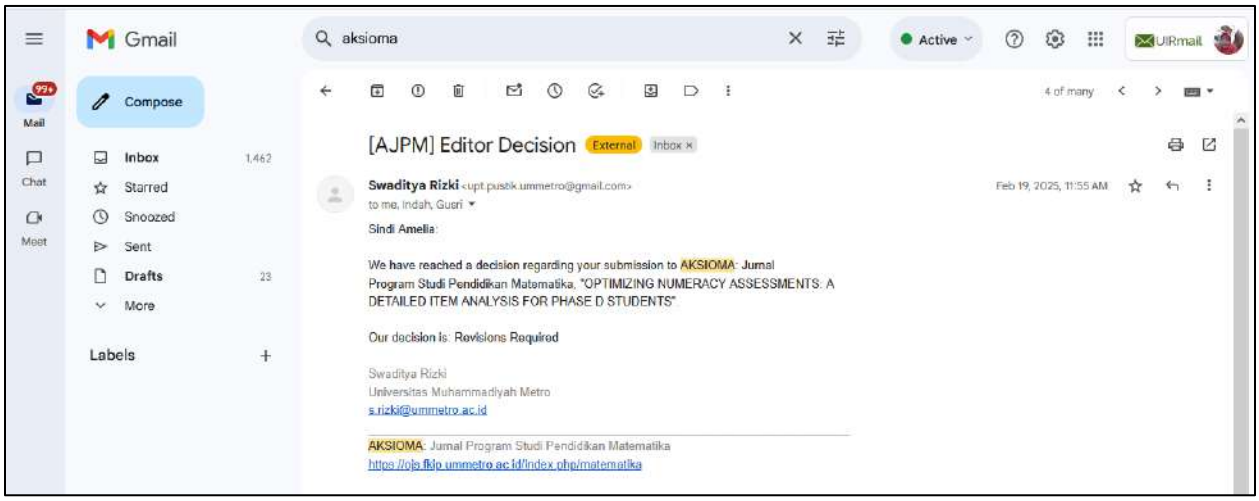
Rizbudiani, A. D., Jaedun, A., Rahim, A., Nurrahman, A. (2021). Rasch Model Item Response Theory (IRT) to Analyze the Quality of Mathematics Final Semester Exam Test on System of Linear Equations in Two Variables (SLETV). Al-Jabar: Jurnal Pendidikan Matematika. 12(2), 399-412. DOI: http://dx.doi.org/10.24042/ajpm.v12i2.9939

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

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

Vahid Aryadoust, Li Ying Ng, and Hiroki Sayama. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing. 38:1, 6-40. https://doi.org/10.1177/0265532220927487

2. Bukti konfirmasi permintaan revisi pertama dan submit hasil revisi yang pertama (19 Februari 2025)



AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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

Received dd Month yy; Received in revised form dd Month yy; Accepted dd Month yy (9pt)

Abstrak

Kemampuan numerasi peserta didik Indonesia yang rendah mestinya menjadi perhatian semua kalangan. Dibandingkan kemampuan literasi, kemampuan numerasi siswa Indonesia lebih memprihatinkan. Untuk itu, praktik-praktik baik yang dapat mendukung peningkatan kemampuan numerasi peserta didik sangat dibutuhkan sebagai bentuk usaha untuk meningkatkan prestasi akademik peserta didik Indonesia. Salah satu praktik baik tersebut adalah dengan rutinnya memberikan soal yang mengukur kemampuan numerasi. Menyusun soal numerasi yang berkualitas, perlu melalui tahapan pengembangan soal yang ilmiah. Pada penelitian sebelumnya, telah dihasilkan instrumen soal numerasi untuk peserta didik fase D yang teruji kevalidan dan kepraktisan melalui serangkaian kegiatan kualitatif (*Self-Evaluation*, *Expert Review*, *One-to-one*, dan *Small Group*). Untuk menyempurnakan kualitas produk soal numerasi untuk peserta didik fase D, maka akan dilanjutkan pada kegiatan kuantitatif, yakni melalui tahapan *Field Test*. Tujuan khusus 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 (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, formative, maupun sumatif untuk mengukur tingkat kemampuan numerasi peserta didik.

Kata kunci: Siswa Fase D; Kemampuan Numerasi; Program Quest.

Abstract

The low numeracy skills of Indonesian students should be a concern for all stakeholders. 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 the regular administration of questions that assess numeracy skills. 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. The specific objective of this study is to analyze 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.*

Comment [HP1]: Gunakan istilah yang sama untuk keseluruhan isi artikel

Comment [HP2]: Untuk apa?
Untuk menganalisis dan memetakan kemampuan numerasi siswa atau untuk melatih siswa dalam mengerjakan soal-soal berbasis kemampuan numerasi

Comment [HP3]: sesuaikan

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

INTRODUCTION

One crucial effort to enhance students' numeracy skills is through regular practice with numeracy questions. 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 analyze 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.

Numeracy Skill

Numeracy is simply viewed dichotomously as testing whether a person can perform basic arithmetic or not (Parnis & Petocz, 2016). Numeracy is not the same as mathematics and is not an alternative to it (Cao Thi et al., 2023). Definitively, numeracy is the ability to develop knowledge and skills by confidently using mathematics to solve practical problems in various aspects of life (Direktorat Sekolah Dasar Kemdikbud, 2021; Jamil & Khusna, 2021). Similarly, mathematical literacy (numeracy) in PISA is defined as an individual's capacity to formulate, use, and interpret mathematics in diverse contexts (Pusat Penilaian Pendidikan, 2019). In other words, numeracy questions tend to refer to PISA-style questions that prioritize reasoning processes over content.

One of the four main policies introduced by the Indonesian Ministry of Education and Culture is to replace the National Examination (UN) with the Minimum Competency Assessment (AKM) and a character survey. The competencies measured in the AKM are literacy and numeracy.

Student Phase D

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

Comment [HP4]: belum ada paparan terkait masalah utama yang mendasari mengapa penel ini perlu untuk dilakukan.

Comment [HP5]: Sesuaikan dengan Template urutan yang harus ditulis dalam PENDAHULUA
1. Perlu sedikit latar belakang umum kajian yan berkaitan dengan judul.
2. *State of the art* (kajian review literatur singk penelitian-penelitian sebelumnya (yang mirip) untuk menjustifikasi ***novelty (Kebaruan)*** artike (harus ada rujukan ke jurnal 10 tahun terakhir);
3. ***Gap analysis*** atau **Pernyataan kesenjangan** (orisinalitas) atau **kebaruan** (*novelty*) penelitia ini dengan penelitian2 sebelumnya yang releva (mirip) atau berdasarkan *state of the art*.
4. Uraikan Permasalahan berdasarkan fakta dan/atau hipotesis (jika ada)
5. Solusi untuk menyelesaikan masalah tesebut
6.hasil yang diharapkan atau tujuan penelitian dalam artikel ini.

Comment [HP6]: Lebih baik paparkan terkait urgensi dari numerasi dan bukan lagi menjelaskar definisi dari numerasi. Lebih penting lagi paparka terkait permasalahan apa yang berkaitan dengan numerasi dan siswa?
Jika kemampuan numerasi siswa rendah, apa penyebabnya?
Apakah dari masalah yang ada memnag yang dibutuhkan adalah pengembangan soal terkait numerasi? Apakah soal yang ada belum cukup ati seperti apa?

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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.

The following table summarizes several studies related to the analysis of numeracy question items:

Comment [HP7]: langsung sebutkan dengan spesifik terkait tabel yang dirujuk

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

No	Research Title / Author Name (Year)	Research Design	Result
1	Development and Validation of Diagnostic Assessment Instrument for Numeracy Skills in 7th Grade / (Burgmanis et al., 2021)	Rasch Model	The diagnostic instrument is suitable for evaluating the numeracy skills of 7th-grade students.
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	Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch / (Nuri Syifa Azzahra et al., 2024)	Rasch Model with Quest Program	Seven out of ten multiple-choice items are appropriate, and all items are valid.
4	Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar / (Tiara Dewi Lestari et al., 2023)	Rasch Model	From the multiple-choice questions tested, it was found that 2 questions fall into the difficult category, 8 questions fall into the moderate category, and 1 question falls into the easy category.

From the four aforementioned studies, item analysis tends to focus on the elementary school level. For Phase D, research subjects are only available in grade 7. 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). ...

Comment [HP8]: justifikasi tujuan penelitian atau hasil yang ingin diperoleh dari peneltian ini

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

METHOD

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 et al., 2023) through the stages of self-evaluation, expert review, one-to-one, and small group, as illustrated in the following diagram:

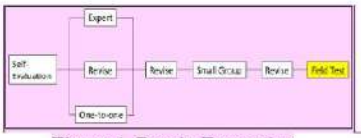


Figure 1. Desain Formative Evaluation

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 analyzing 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 analyze items (Ofianto Ofianto, 2021).

The advantage of this program is its capability to analyze 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 analyze the difficulty level of the Rasch model (Fine Reffiane et al., 2021).

RESULT AND DISCUSSION

The numeracy questions tested quantitatively consist of 30 items, and the percentage distribution of questions based on their components can be seen in the following table:

Comment [HP9]: harap gambar diperbesar/diperjelas

Comment [HP10]: 1.Harus ada Rancangan atau tahapan penelitian secara operasional
2.Subjek, lokasi, dan/atau sampel harus spesifi dan jelas jumlahnya
3.Instrumen Penelitian harus dijelaskan dan te pengumpulan datanya.
4. Teknik Analisis data harus jelas.

[harap dilengkapi]

Comment [HP11]: Sebutkan dengan sepsifik tabel ayng dirujuk

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%

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

Components	Subcomponents	Items	Proportion
	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%

From the table above, 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 estimates, and reliability estimates (Rizbudiani et al., 2021).

Comment [HP12]: Langsung saja "Table 1"

1. Item Validity Estimation

In the Rasch model, the validity of the analyzed items can be observed from the output values of INFIT MNSQ and OUTFIT t (Saryanto et al., 2020). An item item is considered valid if the INFIT MNSQ value falls within the range of "0.5 – 1.5" (Aryadoust et al., 2021) and the OUTFIT t value is "< 2.0" (Abu Bakar et al., 2023; Guo et al., 2020; Muslihin et al., 2022). The validity results of the numeracy items based on the INFIT MNSQ and OUTFIT t values are shown in the following table:

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

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

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

Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation	Item	INFIT MNSQ Value	OUTFIT t Value	Interpretation
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

The table above 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 not decrease data quality, although item 23 specifically can affect reliability scores (Boone et al., 2014). 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 et al., 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 item estimate (Threshold) can determine the difficulty level of the item. The difficulty level of items can be 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 et al., 2023). The analysis results of the difficulty level of numeracy questions for Phase D students can be seen in the following table:

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 *

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00
ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

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

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	-	-

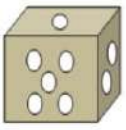
Note: *: not valid

The table above 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 the ~~image below:~~

Comment [HP13]: "Figure 2"

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 2. 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 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) (Bambang Sumintono & Wahyu Widhiarso, 2015). The reliability of item estimate for multiple-choice questions is 0.92 (very good),

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 following table:

Table 6. 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%

From the table above, 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.

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. This also strengthens scientific decisions based on the quantitative analysis of the questions' level of difficulty. Thus, using both approaches (qualitative and quantitative), a collection of numeracy questions that are truly valid, practical, and effective for use in the classroom can be produced.

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

Comment [HP14]: Perlu pembahasan secara mendalam.
1. Authors memberikan argumen terhadap has penelitian yang telah diklaim, ada penjelasan sebab-akibat yang logis dan dirangkai dalam bentuk 'cerita baru' menggunakan kalimat sen
2. Apa temuan dalam penelitian ini.
3. Apa faktor-faktor yang menyebabkan hasil seperti itu
4. Apa kelebihan dan kekurangan dari penelit
5. Bandingkan dengan penelitian yang sebelumnya, apakah ada kesesuaian atau pertentangan dengan hasil penelitian sebelum (**dari state of the art pada PENDAHULUAN**). M penelitian ini sejalan dengan penelitian si A (Tahun), si B (Tahun), si C (Tahun), dst.
6. Harus ada implikasi/dampak/kontribusi hasi penelitian

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

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

difficulty level of the questions, making it easier to make scientific decisions to produce good numeracy questions for Phase D learners.

After being analyzed 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.

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

Abu Bakar, S. N., Mahmood, N., Ismail, A., Hashim, A., Hamsan, N. Y., Nor Hamid, N. I., & Che Yakzam, R. (2023). Validity and Reliability of Competency Analysis Instrument for Cooperative Board Members using Rasch Measurement Model Approach. *International Journal of Academic Research in Business and Social Sciences*, *13*(6), Pages 2408-2419. https://doi.org/10.6007/IJARBSS/v13-i6/17543

Amelia, S., Widiati, I., & Yadrika, G. (2023). PENGEMBANGAN SOAL NUMERASI UNTUK PESERTA DIDIK FASE D. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(3), 3048. https://doi.org/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. https://doi.org/10.1177/0265532220927487

Bambang Sumintono & Wahyu Widhiarso. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Burgmanis, G., France, I., Namsone, D., & Čakāne, L. (2021). *DEVELOPMENT AND VALIDATION OF DIAGNOSTIC ASSESSMENT INSTRUMENT FOR NUMERACY SKILLS IN 7TH GRADE*. 7781–7791. https://doi.org/10.21125/iceri.2021.1747

Cao Thi, H., Le, T. A., Tran Ngoc, B., & Phan Thi Phuong, T. (2023). Factors affecting the numeracy skills of students from mountainous ethnic minority regions in Vietnam: Learners' perspectives. *Cogent Education*, *10*(1), 2202121. https://doi.org/10.1080/2331186X.2023.2202121

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

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

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. https://doi.org/10.21831/reid.v9i1.53514

Direktorat Sekolah Dasar Kemdikbud. (2021). *Modul Literasi Numerasi di Sekolah Dasar*. Jakarta: Kemdikbud.

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. https://doi.org/10.33292/petier.v3i1.51

Fine Reffiane, Sudarmin, Wiyanto, & Sigit Saptono. (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. https://doi.org/10.47750/pegegog.11.04.01

Guo, C., Liu, Y., Hao, S., Xie, L., Xiang, G., Wu, Y., & Li, S. (2020). The Reliability and Validity of the "Activity and Participation" Component in the Brief ICF Core Set for Chronic Obstructive Pulmonary Diseases Based on Rasch Analysis. *International Journal of Chronic Obstructive Pulmonary Disease*, *Volume 15*, 1191–1198. https://doi.org/10.2147/COPD.S249704

Jamil, A. F., & Khusna, A. H. (2021). Pengembankan Asesmen Berorientasi Kontekstual Untuk Meningkatkan Kemampuan Literasi Matematis Dan Numerasi Mahasiswa. *Jurnal Ilmiah Mandala Education*, *7*(4). https://doi.org/10.58258/jime.v7i4.2385

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

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*.

Muslihin, H. Y., Suryana, D., Ahman, A., Suherman, U., & Dahlan, T. H. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch

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

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

Model. *International Journal of Instruction*, *15*(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Nuri Syifa Azzahra, Sri Sumarni, & Himawan Putranta. (2024). Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch. *QuranicEdu: Journal of Islamic Education*, *4*(1), 94. https://doi.org/10.37252/quranicedu.v4i1.681

Ofianto Ofianto. (2021). Analysis Of Instrument Test Of Historical Thinking Skills In Senior High School History Learning With Quest Programs. *Indonesian Journal of History Education*, *6*(2), 184–192.

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. https://doi.org/10.1007/s13384-016-0218-3

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

Pusat Penilaian Pendidikan. (2019). *Pendidikan di IndonesiaL Belajar dari Hasil PISA 2018*. Jakarta: Balitbang Kemdikbud.

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. https://doi.org/10.24042/ajpm.v12i2.9939

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. https://doi.org/10.36681/tused.2020.42

Tiara Dewi Lestari, Ghullam Hamdu, & Erwin Rahayu Saputra. (2023). Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar. *Pendas: Jurnal Ilmu Pendidikan Dasar*, *8*(2), 2489–2503. https://doi.org/10.23969/jp.v8i2.9270

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 dd Month yy; Received in revised form dd Month yy; Accepted dd Month yy (9pt)

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.

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, Experi 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 analyze 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 disiribution (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.*

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 et al., 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 et al., 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 analyze 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

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

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

the government through the Kurikulum Merdeka.

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	Development and Validation of Diagnostic Assessment Instrument for Numeracy Skills in 7th Grade / (Burgmanis et al., 2021)	Rasch Model	The diagnostic instrument is suitable for evaluating the numeracy skills of 7th-grade students.
2	Validation of a Digital Tool for Diagnosing Mathematical Proficiency / (Jumpeng 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	Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch / (Nuri Syifa Azzahra et al., 2024)	Rasch Model with Quest Program	Seven out of ten multiple-choice items are appropriate, and all items are valid.
4	Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar / (Tiara Dewi Lestari et al., 2023)	Rasch Model	From the multiple-choice questions tested, it was found that 2 questions fall into the difficult category, 8 questions fall into the moderate category, and 1 question falls into the easy category.

From the four aforementioned studies, item analysis tends to focus on the elementary school level. For Phase D, research subjects are only available in grade 7. 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 students by applying data analysis techniques with the support of the Quest program.

METHOD

In general, this research constitutes a series of item development studies utilizing a formative evaluation

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 et al., 2023) through the stages of self-evaluation, expert review, one-to-one, and small group, as illustrated in Fig 1.

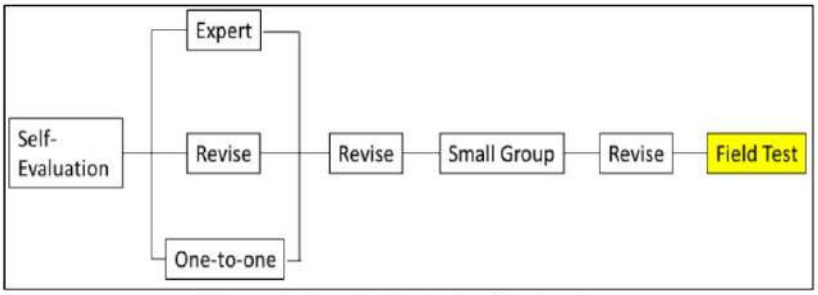


Figure 1. Desain Formative Evaluation

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 analyzing 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 analyze items (Ofianto Ofianto, 2021).

The advantage of this program is its capability to analyze 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 analyze the difficulty level of the Rasch model (Fine Reffiane et al., 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.

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%

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

Components	Subcomponents	Items	Proportion
	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%

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 estimates, and reliability estimates (Rizbudiani et al., 2021) (see Fig. 2).

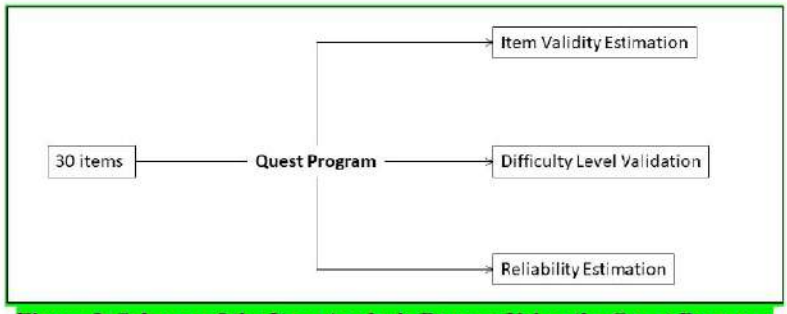


Figure 2. Scheme of the Item Analysis Process Using the Quest Program

In Item Validity Estimation, based on the Rasch Model, the validity of the analyzed items can be assessed using the INFIT MNSQ and OUTFIT t output values (Saryanto et al., 2020). An item is considered valid if the INFIT MNSQ value falls within the range of 0.5 – 1.5 (Aryadoust et al., 2021) and the OUTFIT t value is less than 2.0 (Abu Bakar et al., 2023; Guo et al., 2020; Muslihin et al., 2022).

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)

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

$1 < b \leq 2$ (difficult); 3) $-1 < b \leq 1$ (moderate); 4) $-2 < b \leq -1$ (easy); and 5) $b < -2$ (very easy) (Dewi et al., 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) (Bambang Sumintono & Wahyu Widhiarso, 2015).

RESULT 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 not decrease data quality, although item 23 specifically can affect reliability scores (Boone et al., 2014). The INFIT MNSQ scores for these two items also have a small difference from the validity category threshold (0.5 – 1.5).

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 et al., 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.

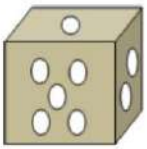
Item number 6 is an essay question with a maximum score of 2. This item is categorized as very difficult, as only 3 out of 34 respondents scored 1, even though none achieved the perfect score. The numeracy component of question number 6 is an essay format, with reasoning as the cognitive level, falling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

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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

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 (Nurhalimah et al., 2022; 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 yet valuable activity for teachers in providing questions as an evaluation tool to their students (Kumar et al., 2021). 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 essential for test development (Quaigrain & Arhin, 2017).

This numeracy test instrument can be further developed to assess the numeracy skills of students in Phase D and analyze 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 analyzed 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.

ACKNOWLEDGEMENTS

Thanks to DRTPM Dikti for improving this article, and DPPM UIR as research sponsor with contract

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

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

number: 486/KONTRAK/P-PT/DPPM-UIR/06-2023.

REFERENCES

Abu Bakar, S. N., Mahmood, N., Ismail, A., Hashim, A., Hamsan, N. Y., Nor Hamid, N. I., & Che Yakzam, R. (2023). Validity and Reliability of Competency Analysis Instrument for Cooperative Board Members using Rasch Measurement Model Approach. *International Journal of Academic Research in Business and Social Sciences*, *13*(6), Pages 2408-2419. https://doi.org/10.6007/IJARBSS/v13-i6/17543

Amelia, S., Widiati, I., & Yadrika, G. (2023). PENGEMBANGAN SOAL NUMERASI UNTUK PESERTA DIDIK FASE D. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(3), 3048. https://doi.org/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. https://doi.org/10.1177/0265532220927487

Bambang Sumintono & Wahyu Widhiarso. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Burgmanis, G., France, I., Namsone, D., & Čakāne, L. (2021). *DEVELOPMENT AND VALIDATION OF DIAGNOSTIC ASSESSMENT INSTRUMENT FOR NUMERACY SKILLS IN 7TH GRADE*. 7781–7791. https://doi.org/10.21125/iceri.2021.1747

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. https://doi.org/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. https://doi.org/10.33292/petier.v3i1.51

Fine Reffiane, Sudarmin, Wiyanto, & Sigit Saptono. (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. https://doi.org/10.47750/pegegog.11.04.01

Guo, C., Liu, Y., Hao, S., Xie, L., Xiang, G., Wu, Y., & Li, S. (2020). The Reliability and

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

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

Validity of the "Activity and Participation" Component in the Brief ICF Core Set for Chronic Obstructive Pulmonary Diseases Based on Rasch Analysis. *International Journal of Chronic Obstructive Pulmonary Disease*, *Volume 15*, 1191–1198. https://doi.org/10.2147/COPD.S249704

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. https://doi.org/10.26803/ijlter.22.1.18

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

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., Faiziyah, N., & 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.

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Muslihin, H. Y., Suryana, D., Universitas Pendidikan Indonesia, Indonesia, dodisuryana@upi.edu, Ahman, A., Universitas Pendidikan Indonesia, Indonesia, ahman@upi.edu, Suherman, U., Universitas Pendidikan Indonesia, Indonesia, umans@upi.edu, Dahlan, T. H., & Universitas Pendidikan Indonesia, Indonesia, tinadahlan_psi@upi.edu. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model. *International Journal of Instruction*, *15*(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). HUBUNGAN ANTARA VALIDITAS ITEM DENGAN DAYA PEMBEDA DAN TINGKAT KESUKARAN SOAL PILIHAN GANDA PAS. *Natural Science Education Research*, *4*(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Nuri Syifa Azzahra, Sri Sumarni, & Himawan Putranta. (2024). Analisis Validitas dan Realibilitas

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

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

Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch. *QuranicEdu: Journal of Islamic Education*, *4*(1), 94. https://doi.org/10.37252/quranicedu.v4i1.681

Ofianto Ofianto. (2021). Analysis Of Instrument Test Of Historical Thinking Skills In Senior High School History Learning With Quest Programs. *Indonesian Journal of History Education*, *6*(2), 184–192.

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. https://doi.org/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. https://doi.org/10.1080/2331186X.2017.1301013

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. https://doi.org/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. https://doi.org/10.46328/ijemst.3513

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. https://doi.org/10.36681/tused.2020.42

Tiara Dewi Lestari, Ghullam Hamdu, & Erwin Rahayu Saputra. (2023). Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar. *Pendas: Jurnal Ilmu Pendidikan Dasar*, *8*(2), 2489–2503. https://doi.org/10.23969/jp.v8i2.9270

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*,

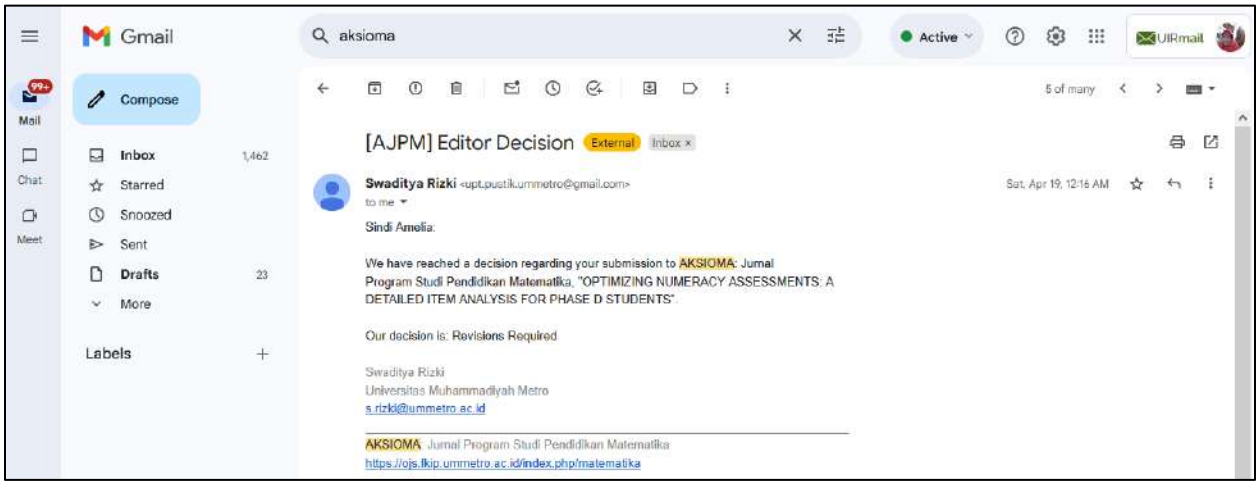
ISSN 2089-8703 (Print)
ISSN 2442-5419 (Online)

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

43(13), 2185–2205. https://doi.org/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. https://doi.org/10.31893/multiscience.2025410

3. Bukti konfirmasi permintaan revisi kedua dan submit revisi yang kedua (19 April 2025)



AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 dd Month yy; Received in revised form dd Month yy; Accepted dd Month yy (9pt)

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.

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 analyze 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.*

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

INTRODUCTION

Numeracy is simply viewed dichotomously as testing whether a person can perform basic aritlunetic 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 et al., 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 et al., 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 analyze 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

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

the government through the Kurikulum Merdeka.

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	Development and Validation of Diagnostic Assessment Instrument for Numeracy Skills in 7th Grade / (Burgmanis et al., 2021)	Rasch Model	The diagnostic instrument is suitable for evaluating the numeracy skills of 7th-grade students.
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	Analisis Validitas dan Realibilitas Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch / (Nuri Syifa Azzahra et al., 2024)	Rasch Model with Quest Program	Seven out of ten multiple-choice items are appropriate, and all items are valid.
4	Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar / (Tiara Dewi Lestari et al., 2023)	Rasch Model	From the multiple-choice questions tested, it was found that 2 questions fall into the difficult category, 8 questions fall into the moderate category, and 1 question falls into the easy category.

From the four aforementioned studies, item analysis tends to focus on the elementary school level. For Phase D, research subjects are only available in grade 7. 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 students by applying data analysis techniques with the support of the Quest program.

METHOD

In general, this research constitutes a series of item development studies utilizing a formative evaluation

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 et al., 2023) through the stages of self-evaluation, expert review, one-to-one, and small group, as illustrated in Fig 1.

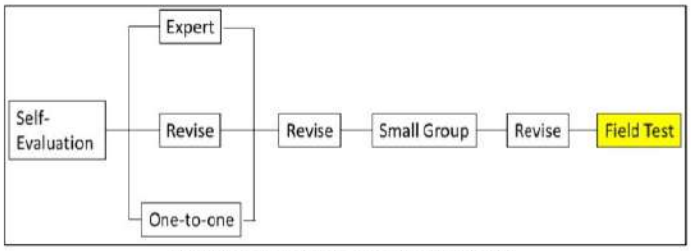


Figure 1. Desain Formative Evaluation

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 analyzing 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 analyze items (Ofianto Ofianto, 2021).

The advantage of this program is its capability to analyze 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 analyze the difficulty level of the Rasch model (Fine Reffiane et al., 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.

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%

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

Components	Subcomponents	Items	Proportion
	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%

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 estimates, and reliability estimates (Rizbudiani et al., 2021) (see Fig. 2).

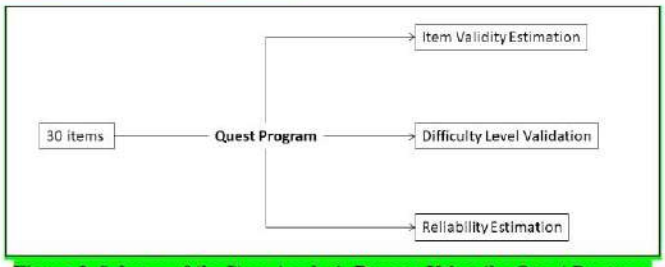


Figure 2. Scheme of the Item Analysis Process Using the Quest Program

In Item Validity Estimation, based on the Rasch Model, the validity of the analyzed items can be assessed using the INFIT MNSQ and OUTFIT t output values (Saryanto et al., 2020). An item is considered valid if the INFIT MNSQ value falls within the range of 0.5 – 1.5 (Aryadoust et al., 2021) and the OUTFIT t value is less than 2.0 (Abu Bakar et al., 2023; Guo et al., 2020; Muslihin et al., 2022).

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)

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

$1 < b \leq 2$ (difficult); 3) $-1 < b \leq 1$ (moderate); 4) $-2 < b \leq -1$ (easy); and 5) $b < -2$ (very easy) (Dewi et al., 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) (Bambang Sumintono & Wahyu Widhiarso, 2015).

RESULT 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 not decrease data quality, although item 23 specifically can affect reliability scores (Boone et al., 2014). The INFIT MNSQ scores for these two items also have a small difference from the validity category threshold (0.5 – 1.5).

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

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

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 et al., 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 0, No. 0, 20xx, 00-00

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

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

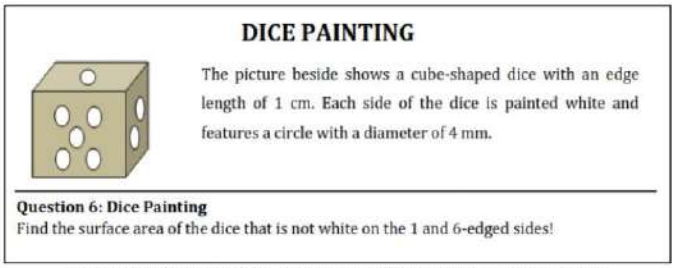


Figure 3. Question Item Number 6 (Very Difficult Category)

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

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 (Nurhalimah et al., 2022; 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 yet valuable activity for teachers in providing questions as an evaluation tool to their students (Kumar et al., 2021). 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 essential for test development (Quaigrain & Arhin, 2017).

This numeracy test instrument can be further developed to assess the numeracy skills of students in Phase D and analyze 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 analyzed 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.

ACKNOWLEDGEMENTS

Thanks to DRTPM Dikti for improving this article, and DPPM UIR as research sponsor with contract

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

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

number: 486/KONTRAK/P-PT/DPPM-UIR/06-2023.

REFERENCES

Abu Bakar, S. N., Mahmood, N., Ismail, A., Hashim, A., Hamsan, N. Y., Nor Hamid, N. I., & Che Yakzam, R. (2023). Validity and Reliability of Competency Analysis Instrument for Cooperative Board Members using Rasch Measurement Model Approach. *International Journal of Academic Research in Business and Social Sciences*, *13*(6), Pages 2408-2419. https://doi.org/10.6007/IJARBSS/v13-i6/17543

Amelia, S., Widiati, I., & Yadrika, G. (2023). PENGEMBANGAN SOAL NUMERASI UNTUK PESERTA DIDIK FASE D. *AKSIOMA: Jurnal Program Studi Pendidikan Matematika*, *12*(3), 3048. https://doi.org/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. https://doi.org/10.1177/0265532220927487

Bambang Sumintono & Wahyu Widhiarso. (2015). *Aplikasi Pemodelan Rasch pada Assessment Pendidikan*. Trim Komunikata.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Burgmanis, G., France, I., Namsone, D., & Čakāne, L. (2021). *DEVELOPMENT AND VALIDATION OF DIAGNOSTIC ASSESSMENT INSTRUMENT FOR NUMERACY SKILLS IN 7TH GRADE*. 7781–7791. https://doi.org/10.21125/iceri.2021.1747

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. https://doi.org/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. https://doi.org/10.33292/petier.v3i1.51

Fine Reffiane, Sudarmin, Wiyanto, & Sigit Saptono. (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. https://doi.org/10.47750/pegegog.11.04.01

Guo, C., Liu, Y., Hao, S., Xie, L., Xiang, G., Wu, Y., & Li, S. (2020). The Reliability and

Comment [A1]: Ada beberapa sumber yg metadatanya belum lengkap seperti tidak ada nama jurnal, volume, nomor, halaman, doi, dsb

Comment [A2]: Tidak relevan dengan penelitian anda

~~Guo, C., Liu, Y., Hao, S., Xie, L., Xiang, G., Wu, Y., & Li, S. (2020). The Reliability and Validity of the "Activity and Participation" Component in the Brief ICF Core Set for Chronic Obstructive Pulmonary Diseases Based on Rasch Analysis. *International Journal of Chronic Obstructive Pulmonary Disease*, *Volume 15*, 1191–1198. https://doi.org/10.2147/COPD.S249700~~

Comment [A3]: Sumber belum memenuhi standar ini

1. Jumlah referensi minimal 20 (harus relevan dengan judul artikel anda), lebih dari 80% referensi harus berasal dari sumber primer (primer > 80% atau >= 81%, sumber primer yaitu jurnal penelitian, skripsi/thesis/disertasi). Lebih utama dari Jurnal semua.
2. Kutipan dari buku teks/teori maksimal 10% dari total referensi.
3. Sumber diutamakan berasal dari Jurnal Internasional Bereputasi (SCOPUS, WoS, dsb) Jurnal Nasional Terakreditasi Sinta 1-3 https://sinta.kemdikbud.go.id/journals.
4. Referensi 10 Tahun terakhir, Lebih utama jika lebih banyak referensi 5 tahun terakhir
5. Format daftar pustaka Gunakan APA Style edisi ke-6.
6. Semua kutipan dan daftar pustaka wajib menggunakan reference manager : mendeley, zotero, atau ms word reference, dll [WAJIB]
7. Cek kelengkapan metadata seperti nama penulis, tahun, judul, nama jurnal, volume, nomor, halaman, doi wajib ada
8. Referensi harus relevan dengan judul penelitian anda dan dalam bidang pendidikan matematika

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

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

Validity of the "Activity and Participation" Component in the Brief ICF Core Set for Chronic Obstructive Pulmonary Diseases Based on Rasch Analysis. *International Journal of Chronic Obstructive Pulmonary Disease*, *Volume 15*, 1191–1198. https://doi.org/10.2147/COPD.S249704

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. https://doi.org/10.26803/ijlter.22.1.18

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

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., Faiziyah, N., & 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.

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/10.1016/j.mjafi.2020.11.007

Muslihin, H. Y., Suryana, D., Universitas Pendidikan Indonesia, Indonesia, dodisuryana@upi.edu, Ahman, A., Universitas Pendidikan Indonesia, Indonesia, ahman@upi.edu, Suherman, U., Universitas Pendidikan Indonesia, Indonesia, unans@upi.edu, Dahlan, T. H., & Universitas Pendidikan Indonesia, Indonesia, tinadahlan_psi@upi.edu. (2022). Analysis of the Reliability and Validity of the Self-Determination Questionnaire Using Rasch Model. *International Journal of Instruction*, *15*(2), 207–222. https://doi.org/10.29333/iji.2022.15212a

Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). HUBUNGAN ANTARA VALIDITAS ITEM DENGAN DAYA PEMBEDA DAN TINGKAT KESUKARAN SOAL PILIHAN GANDA PAS. *Natural Science Education Research*, *4*(3), 249–257. https://doi.org/10.21107/nser.v4i3.8682

Nuri Syifa Azzahra, Sri Sumarni, & Himawan Putranta. (2024). Analisis Validitas dan Realibilitas

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

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

~~Kualitas Soal Pilihan Ganda Asesmen Kompetensi Minimum (AKM) Mata Pelajaran Pendidikan Agama Islam Menggunakan Pendekatan Model Rasch. *QuranicEdu: Journal of Islamic Education*, *4*(1), 94. https://doi.org/10.37252/quranicedu.v4i1.681~~

~~Ofianto Ofianto. (2021). Analysis Of Instrument Test Of Historical Thinking Skills In Senior High School History Learning With Quest Programs. *Indonesian Journal of History Education*, *6*(2), 184–192.~~

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. https://doi.org/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. https://doi.org/10.1080/2331186X.2017.1301013

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. https://doi.org/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. https://doi.org/10.46328/ijemst.3513

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. https://doi.org/10.36681/tused.2020.42

Tiara Dewi Lestari, Ghullam Hamdu, & Erwin Rahayu Saputra. (2023). Analisis Soal Literasi Numerasi Menggunakan Pemodelan Rasch Konteks Pemanasan Global Berbasis ESD untuk Sekolah Dasar. *Pendas: Jurnal Ilmu Pendidikan Dasar*, *8*(2), 2489–2503. https://doi.org/10.23969/jp.v8i2.9270

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*,

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

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

43(13), 2185–2205. https://doi.org/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. https://doi.org/10.31893/multiscience.2025410

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

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 dd Month yy; Received in revised form dd Month yy; Accepted dd Month yy (9pt)

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.

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.*

AKSIOMA: Jurnal Program Studi Pendidikan Matematika
Volume 0, No. 0, 20xx, 00-00

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

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

INTRODUCTION

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

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

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

the government through the Kurikulum Merdeka.

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 et al. 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 / (Jumpeng 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 et al. 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

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

METHOD

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, and Yadrika 2023) through the stages of self-evaluation, expert review, one-to-one, and small group, as illustrated in Fig 1.

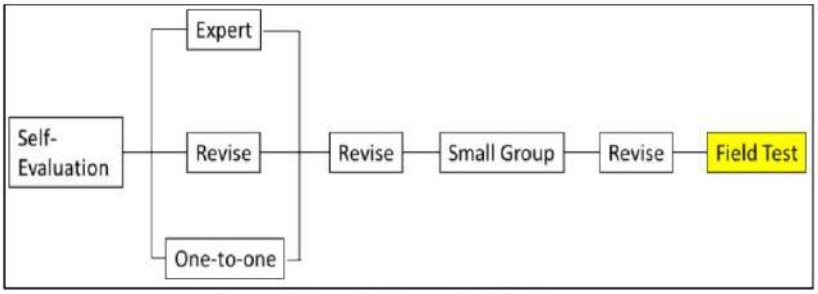


Figure 1. Desain Formative Evaluation

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 analyzing 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 et al. 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.

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,	33%

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

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

Components	Subcomponents	Items	Proportion
		23	
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%

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 estimates, and reliability estimates (Rizbudiani et al. 2021) (see Fig. 2).

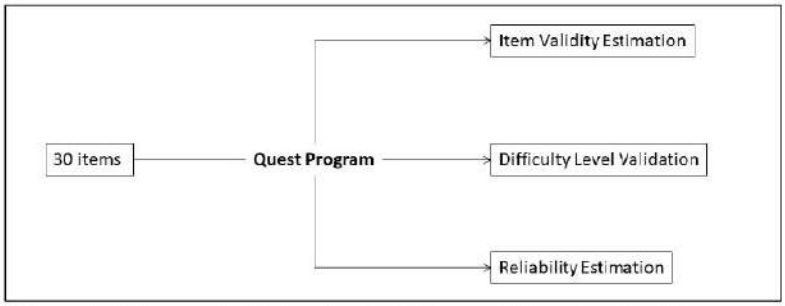


Figure 2. Scheme of the Item Analysis Process Using the Quest Program

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 et al. 2020). An item is considered valid if the INFIT MNSQ value falls within the range of 0.5 – 1.5 (Aryadoust, Ng, and Sayama 2021) and

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

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

the OUTFIT t value is less than 2.0 (Alhadabi and Aldhafri 2021; Rusyid et al. 2024; Sukarelawan et al. 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, and 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 and Widhiarso 2015).

RESULT 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.

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

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

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 not decrease data quality, although item 23 specifically can affect reliability scores (Adi et al. 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 et al. 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,

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

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

and using a scientific context, as shown in Fig 3.

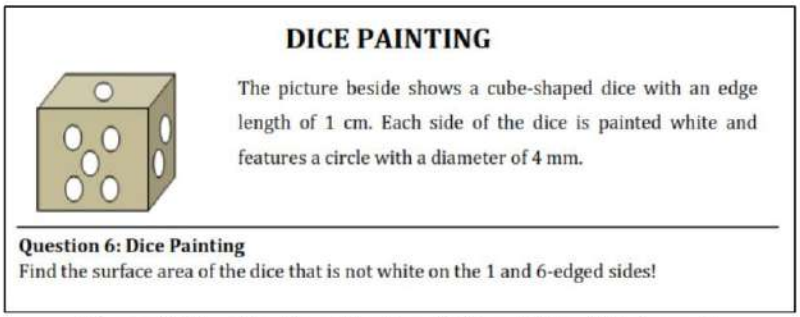


Figure 3. Question Item Number 6 (Very Difficult Category)

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

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, and Cormier 2019; Van Vo and 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, and Eom 2021), yet it can show which items are useful (Asempapa and 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 et al. 2024; Quaigrain and 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.

ACKNOWLEDGEMENTS

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

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

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, Nur Romdlon Maslahul, Hidar Amaruddin, Habib Maulana Maslahul Adi, and Isna Laili Qurroti A'yun. 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):103–13. doi: https://doi.org/10.15294/jere.v11i2.58835.

Alhadabi, Amal, and Said Aldhafri. 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):957–73. doi: 10.12973/eu-jer.10.2.957.

Alsina, Angel, Antonio Maurandi, Elvira Ferre, and Claudia Coronata. 2021. 'Validating an Instrument to Evaluate the Teaching of Mathematics Through Processes'. *International Journal of Science and Mathematics Education* 19(3):559–77. doi: 10.1007/s10763-020-10064-y.

Amelia, Sindi, Indah Widiati, and Gusri Yadrika. 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, Vahid, Li Ying Ng, and Hiroki Sayama. 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, Reuben S., and Doris Lee. 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, Shiau-Wei, Chee-Kit Looi, Weng Kin Ho, Wendy Huang, Peter Seow, and Longkai Wu. 2021. 'Learning Number Patterns through Computational Thinking Activities: A Rasch Model Analysis'. *Heliyon* 7(9):e07922. doi: 10.1016/j.heliyon.2021.e07922.

Dewi, Henda Harmantia, Siti Maftuhah Damio, and Sukarno Sukarno. 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, Muhammad, Mohammad Archi Maulyda, Ida Ermiana, Vivi Rachmatul Hidayati, and Arif Widodo. 2020. 'Validity and Reliability of Cognitive Tests Study and Development of

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

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

Elementary Curriculum Using Rasch Model'. *Psychology, Evaluation, and Technology in Educational Research* 3(1):26–33. doi: 10.33292/petier.v3i1.51.

Ing, Marsha, Karl W. Kosko, Cindy Jong, and Jeffrey C. Shih. 2024. 'Validity Evidence of the Use of Quantitative Measures of Students in Elementary Mathematics Education'. *School Science and Mathematics* 124(6):411–23. doi: 10.1111/ssm.12660.

Ismawati, Esti, Hersulastuti Hersulastuti, Indiyah Prana Amertawengrum, and Kun Andyan Anindita. 2023. 'Portrait of Education in Indonesia: Learning from PISA Results 2015 to Present'. *International Journal of Learning, Teaching and Educational Research* 22(1):321–40. doi: 10.26803/ijlter.22.1.18.

Isnani, Isnani, Wikan Budi Utami, Purwo Susongko, and Herani Tri Lestiani. 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, Putcharee, Metta Marwiang, Samruan Chinjunthuk, Prapawadee Suwannatrai, Kanokporn Chanayota, Kissadapan Pongboriboon, Keow Ngang Tang, and Mark Wilson. 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, Adnan, Okan Bulut, and Damien C. Cormier. 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, Muhammad Noor, Faizatur Rofi'ah, Naufal Ishartono, Mohamad Waluyo, Swasti Maharani, Annisa Swastika, Nuqthy Faiziyah, and Christina Kartika Sari. 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, Seock-Ho, Allan S. Cohen, and Hyo Jin Eom. 2021. 'A Note on the Three Methods of Item Analysis'. *Behaviormetrika* 48(2):345–67. doi: 10.1007/s41237-021-00131-1.

Parnis, Amanda J., and Peter Petocz. 2016. 'Secondary School Students' Attitudes towards Numeracy: An Australian Investigation Based on the National Assessment Program—

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

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

Literacy and Numeracy (NAPLAN)'. *The Australian Educational Researcher* 43(5):551–66. doi: 10.1007/s13384-016-0218-3.

Pusat Asesmen dan Pembelajaran. 2020. 'Desain Pengembangan Soal AKM'.

Quaigrain, Kennedy, and Ato Kwamina Arhin. 2017. 'Using Reliability and Item Analysis to Evaluate a Teacher-Developed Test in Educational Measurement and Evaluation' edited by S. King Fai Hui. *Cogent Education* 4(1):1301013. doi: 10.1080/2331186X.2017.1301013.

Reffiane, Fine, Sudarmin, Wiyanto, and Sigit Saptono. 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.

Rizbudiani, Adilla Desy, Amat Jaedun, Abdul Rahim, and Arief Nurrahman. 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, Rosnelli, and Pitra Ashrin Ristiana. 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):946–63. doi: 10.46328/ijemst.3513.

Rusyid, Husnul Khatimah, Didi Suryadi, Tatang Herman, Mazlini Adnan, Ahmad Lutfi, and Ahmad Mukhibin. 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, Saryanto, Rudy Sumiharsono, Syahrul Ramadhan, and Edy Suprapto. 2020. 'The Analysis of Instrument Quality to Measure theStudents Higher Order Thinking Skill in Physics Learning'. *Turkish Journal of Science Education* 17(4):520–27. doi: 10.36681/tused.2020.42.

Sinaga, Nurul Afni. 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. Irma, Jumadi Jumadi, Heru Kuswanto, and M. Anas Thohir. 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–

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

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

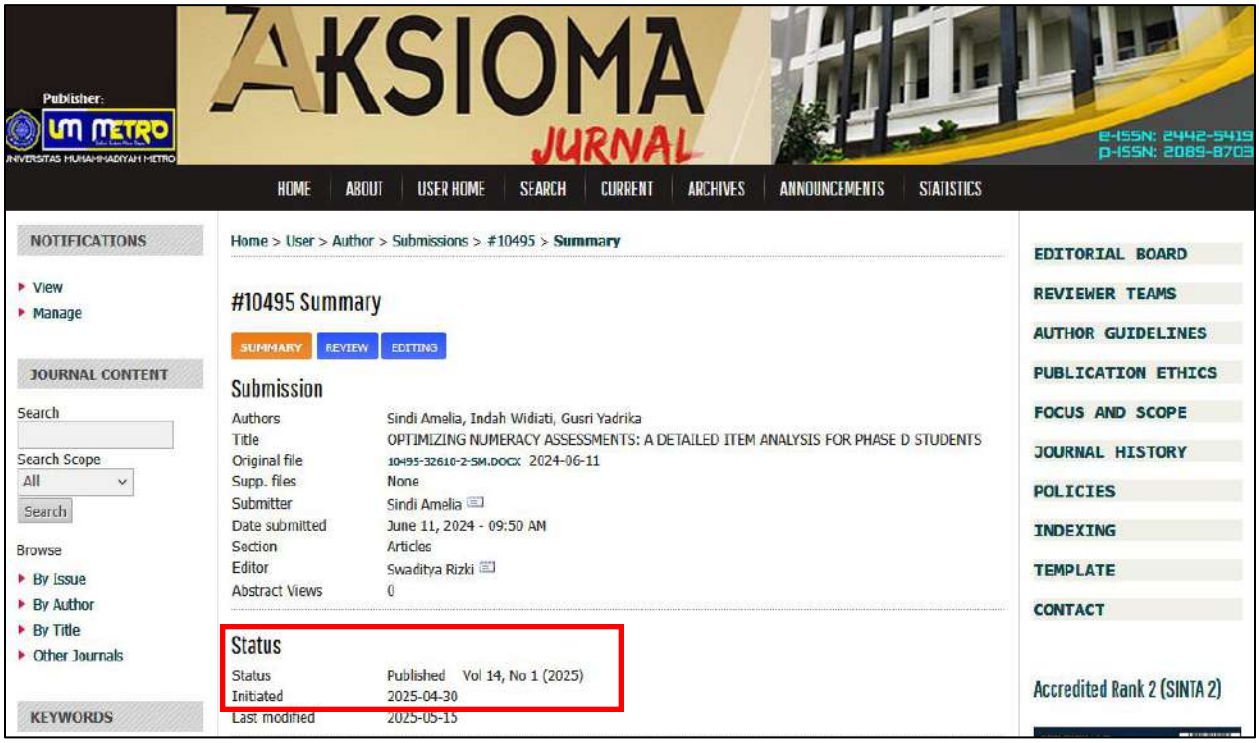
october–2021):2133–44. doi: 10.12973/eu-jer.10.4.2133.

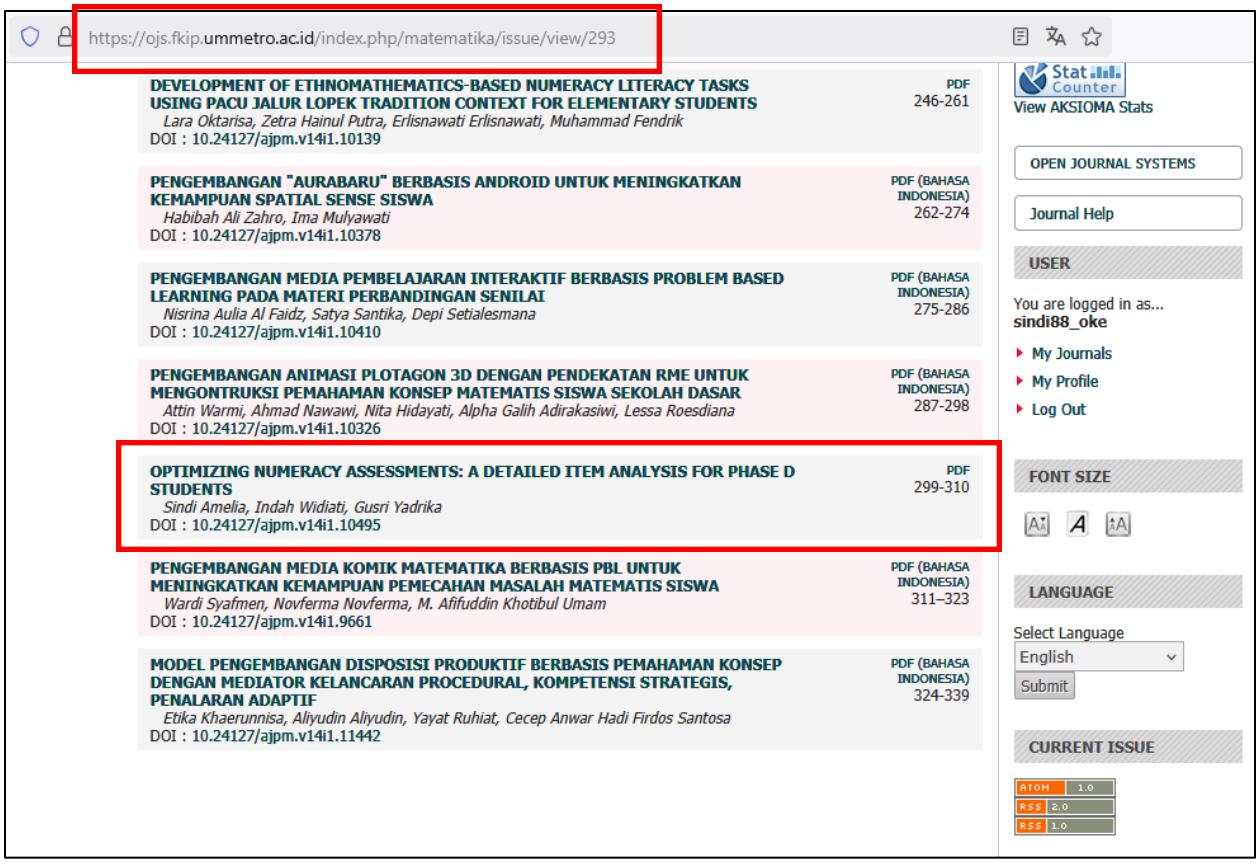
Sumintono, Bambang, and Wahyu Widhiarso. 2015. *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan*. Cimahi, Indonesia: Trim Komunikata.

Van Vo, De, and Benő Csapó. 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, Via, Dian Kusmaharti, and Imas Srinana Wardani. 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.

4. Bukti konfirmasi artikel accepted dan published di web jurnal (30 April 2025)




https://ojs.fkip.ummetro.ac.id/index.php/matematika/issue/view/293
DEVELOPMENT OF ETHNOMATHEMATICS-BASED NUMERACY LITERACY TASKS USING PACU JALUR LOPEK TRADITION CONTEXT FOR ELEMENTARY STUDENTS
Lara Oktarisa, Zetra Hainul Putra, Erlisnawati Erlisnawati, Muhammad Fendrik
DOI : 10.24127/ajpm.v14i1.10139
PDF 246-261
PENGEMBANGAN "AURABARU" BERBASIS ANDROID UNTUK MENINGKATKAN KEMAMPUAN SPATIAL SENSE SISWA
Habibah Ali Zahro, Ima Mulyawati
DOI : 10.24127/ajpm.v14i1.10378
PDF (BAHASA INDONESIA) 262-274
PENGEMBANGAN MEDIA PEMBELAJARAN INTERAKTIF BERBASIS PROBLEM BASED LEARNING PADA MATERI PERBANDINGAN SENILAI
Nisrina Aulia Al Faidz, Satya Santika, Depi Setialesmana
DOI : 10.24127/ajpm.v14i1.10410
PDF (BAHASA INDONESIA) 275-286
PENGEMBANGAN ANIMASI PLOTAGON 3D DENGAN PENDEKATAN RME UNTUK MENGONTRUKSI PEMAHAMAN KONSEP MATEMATIS SISWA SEKOLAH DASAR
Attin Warmi, Ahmad Nawawi, Nita Hidayati, Alpha Galih Adirakasiwi, Lessa Roesdiana
DOI : 10.24127/ajpm.v14i1.10326
PDF (BAHASA INDONESIA) 287-298
OPTIMIZING NUMERACY ASSESSMENTS: A DETAILED ITEM ANALYSIS FOR PHASE D STUDENTS
Sindi Amelia, Indah Widiati, Gusri Yadrika
DOI : 10.24127/ajpm.v14i1.10495
PDF 299-310
PENGEMBANGAN MEDIA KOMIK MATEMATIKA BERBASIS PBL UNTUK MENINGKATKAN KEMAMPUAN PEMECAHAN MASALAH MATEMATIS SISWA
Wardi Syafmen, Novferma Novferma, M. Afifuddin Khotibul Umam
DOI : 10.24127/ajpm.v14i1.9661
PDF (BAHASA INDONESIA) 311–323
MODEL PENGEMBANGAN DISPOSISI PRODUKTIF BERBASIS PEMAHAMAN KONSEP DENGAN MEDIATOR KELANCARAN PROCEDURAL, KOMPETENSI STRATEGIS, PENALARAN ADAPTIF
Etika Khaerunnisa, Aliyudin Aliyudin, Yayat Ruhiat, Cecep Anwar Hadi Firdos Santosa
DOI : 10.24127/ajpm.v14i1.11442
PDF (BAHASA INDONESIA) 324-339
View AKSIOMA Stats
OPEN JOURNAL SYSTEMS
Journal Help
USER
You are logged in as...
sindi88_oke
My Journals
My Profile
Log Out
FONT SIZE
LANGUAGE
Select Language
English
Submit
CURRENT ISSUE


https://ojs.fkip.ummetro.ac.id/index.php/matematika/article/view/10495
AKSIOMA
Jurnal Program Studi Pendidikan Matematika
Publisher:
UM METRO
UNIVERSITAS MUHAMMADIYAH METRO
AKSIOMA JURNAL
e-ISSN: 2442-5419
p-ISSN: 2089-8703
HOME ABOUT USER HOME SEARCH CURRENT ARCHIVES ANNOUNCEMENTS STATISTICS
NOTIFICATIONS
View
Manage
JOURNAL CONTENT
Search
Search Scope
All
Search
Browse
By Issue
By Author
By Title
Other Journals
Home > Vol 14, No 1 (2025) > Amelia
OPTIMIZING NUMERACY ASSESSMENTS: A DETAILED ITEM ANALYSIS FOR PHASE D STUDENTS
Sindi Amelia(1*), Indah Widiati(2), Gusri Yadrika(3),
(1) Universitas Islam Riau
(2) Universitas Islam Riau
(3) Universitas Riau
(*) Corresponding Author
Abstract
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
EDITORIAL BOARD
REVIEWER TEAMS
AUTHOR GUIDELINES
PUBLICATION ETHICS
FOCUS AND SCOPE
JOURNAL HISTORY
POLICIES
INDEXING
TEMPLATE
CONTACT