JURNAL ABSIS
Jurnal Pendidikan Matematika dan Matematika
Vol. 9 No. 1, April 2026

EVALUATING THE RELIABILITY OF MATHEMATICAL LITERACY ASSESSMENT FOR PRE-SERVICE TEACHERS: A GENERALIZABILITY THEORY APPROACH

Lusi Eka Afri[1], Nurrahmawati[2], Arcat[3]
[1-3] Universitas Pasir Pengaraian, Indonesia
lusiekaafri13@gmail.com

ABSTRACT Reliable assessment of pre-service teachers' mathematical literacy skills is crucial to ensure their readiness to teach. However, conventional reliability indices often fail to disentangle various sources of measurement error simultaneously. This study applies Generalizability Theory (G-Theory) to evaluate and diagnose the reliability of a mathematics literacy instrument. A two-facet nested design, denoted as $p \times (i{:}r)$, was employed involving 30 undergraduate students (p) with six open ended essay item (i) evaluated by three independent raters (r). A G-study was conducted to estimated variance components, revealing that the largest source of variance originated from raters (92.8%). The initial analysis yielded d generalizability coefficient of 0.27 and dependability coefficient of 0.01, indicating low reliability due to substantial rater inconsistency. A D-study was conducted to evaluate alternative designs, however the results showed that even by increasing the number of items and raters, the G-coefficient 0.38 above the threshold of 0.80. These findings highlight that rater effects pose a major threat to measurement stability and suggest that the current instrument requires fundamental refinement of its scoring rubrics and procedures. This study serves as a critical diagnostic for teacher education programs, providing an empirical basis for optimizing assessment design and institutionalizing rater calibration to improve the evaluation of future educators.

Keywords: assessment reliability, generalizability theory, mathematical literacy, pre-service teachers

ABSTRAK Penilaian yang andal terhadap kemampuan literasi matematis calon guru sangat penting untuk memastikan kesiapan mereka dalam mengajar. Namun, indeks reliabilitas konvensional sering kali tidak mampu memisahkan berbagai sumber kesalahan pengukuran secara simultan. Penelitian ini menerapkan *Generalizability Theory* (G-Theory) untuk mengevaluasi dan mendiagnosis reliabilitas instrumen literasi matematika. Desain bersarang dua faset, yang dinotasikan sebagai $p \times (i{:}r)$, digunakan dengan melibatkan 30 mahasiswa calon guru (p), enam butir soal esai terbuka (i), dan tiga penilai independen $(r$. G-study dilakukan untuk mengestimasi komponen varians, yang menunjukkan bahwa sumber varians terbesar berasal dari penilai, yaitu sebesar 92,8%. Analisis awal menghasilkan koefisien generalizability sebesar 0,27 dan koefisien dependability sebesar 0,01, yang menunjukkan

reliabilitas rendah akibat inkonsistensi penilai yang substansial. D-study dilakukan untuk mengevaluasi alternatif desain pengukuran; namun, hasilnya menunjukkan bahwa meskipun jumlah butir soal dan penilai ditingkatkan, koefisien G hanya mencapai 0,38 dan masih berada di bawah ambang batas 0,80. Temuan ini menunjukkan bahwa efek penilai menjadi ancaman utama terhadap stabilitas pengukuran dan mengindikasikan bahwa instrumen saat ini memerlukan penyempurnaan mendasar pada rubrik penskoran dan prosedur penilaian. Penelitian ini berperan sebagai diagnosis kritis bagi program pendidikan guru dengan memberikan dasar empiris untuk mengoptimalkan desain asesmen dan melembagakan kalibrasi penilai guna meningkatkan evaluasi calon pendidik di masa depan.

Kata-kata kunci: reliabilitas asesmen, generalizability theory, literasi matematis, calon guru

INTRODUCTION

Mathematical literacy is the ability to understand, use, and interpret mathematical concepts and procedures in various everyday life contexts. According to PISA (OECD, 2017) mathematical literacy is defined as an individual's ability to formulate, use, and interpret mathematics in a variety of real-world contexts. This includes skills such as using mathematical concepts, procedures, facts, and tools to describe and predict phenomena, as well as mathematical reasoning abilities. For instance, mathematical literacy can help individuals understand and manage their personal finances, make informed purchasing decisions, and solve problems in their workplaces.

In the context of education, mathematical literacy is crucial not only for helping students master basic mathematical concepts but also for enriching their understanding with conceptual comprehension of how mathematics is applied in real-world situations. Furthermore, mathematical literacy also plays a role in developing critical thinking skills, enabling students to identify, formulate, and solve problems using mathematical approaches, as well as argue logically and communicatively. As (Ojose, 2011) notes, mathematical literacy is the knowledge to understand and apply basic mathematics in everyday life. This highlights the role of mathematical literacy in preparing students to face everyday challenges effectively and efficiently.

In Indonesia, the development of mathematical literacy is reflected in the framework of the Minimum Competency Assessment (AKM). This national framework is explicitly designed to align with international benchmarks, adopting the literacy-based evaluation principles popularized by PISA while specifically referring to the cognitive standard developed by the International Association for the Evaluation of Education Achievement (IEA) through the TIMSS framework (Wijaya & Dewayani, 2021). by synthesizing these global standards, the AKM categorized mathematical literacy specifically numeracy into three cognitive levels, namely knowing, applying, and reasoning (Mullis, I. V. S. & Martin, 2017). Consequently, the instruments used to assess mathematical literacy skills cover these three cognitive levels. The knowing level consists of aspects such as recalling, identifying facts, and performing algebraic procedures effectively. The applying level consists of aspects such as applying

https://doi.org/10.30606/absis.v9i1.3585

strategies and operations and interpreting problem solutions. Meanwhile, the reasoning level consists of aspects such as analyzing, drawing conclusions, and providing mathematical arguments.

In response to 21st-century developments, teachers with robust mathematical literacy are essential to guide students in solving real-word problems and reducing educational disparities, and building a foundation for an innovative future (Whitney-smith et al., 2022). However, the condition of low numeracy skills among pre-service teachers (Ayuningtyas, N. & Sukriyah, 2020; Basri et al., 2021) poses a significant challenge for higher education institutions. Addressing this requires not only curriculum strengthening but also effective evaluation instrument for pre-service teachers that provide consistent and meaningful results (Allen & Yen, 2001).

In this regard, mathematical literacy assessment instruments are fundamental in determining whether pre-service teachers possess the necessary qualifications to guide student in understanding and applying mathematical concepts. For these prospective educators, such assessment serve as a solid foundation for applying and integrating their knowledge in the context learning. Thus, ensuring the reliability of these instruments is paramount to guarantee that pre-service teachers are truly prepared to face future learning challenges and meet professional standards.

In addition to the technical aspects of the instrument, the quality of these assessment results is also heavily influenced by the assessment procedures used. Assessment procedures refer to the steps or methods used to measure, evaluate, and assess a particular aspect (Bulková et al., 2022). The instrument used in this study was developed based on mathematical literacy indicators that refer to three cognitive levels in AKM, namely knowing, applying, and reasoning. Given that the items are open-ended essays (constructed-response tasks) required students to explain reasoning, interpret data, the assessment cannot be conducted singularly. Therefore, multiple independent raters were involved to maintain objectivity and mitigate the inherent subjectivity od essay type evaluation.

Despite the need for precision, most existing evaluations in Indonesia particularly within teacher education program like the one at Pasir Pengaraian University still rely on Classical Test Theory (CTT). CTT is limited by its inability to assess the effect of multiple sources of measurement error simultaneously, as it typically treats error as a single, undifferentiated entity and provides only a single reliability index, such as Cronbach's alpha. Vispoel et al., (2018), single occasion indices like alpha often overestimate score consistency because they fail to disentangle specific factor and transient errors from the true universe score. In the context of open-ended essay tasks, CTT is insufficient as it cannot distinguish whether score inconsistencies arise from students' actual abilities, item complexity, or the inherent subjectivity of the lecturers acting as raters.

The novelty of this study lies in the application of Generalizability Theory (G-Theory) to simultaneously dissect multiple identifiable sources of error within a

mathematical literacy instrument aligned with AKM standards. G-Theory allows for the evaluation of multiple sources of error simultaneously, determining the extent to which assessment result can be generalized across various measurement conditions (Tan, 2023). Unlike previous studies that merely report a general reliability coefficient, this research provides a unique methodological contribution through Decision Study (D-Study) analysis. D-Study is used to make optimal decisions about optimal assessment design, such as the number of items or raters required to achieve a certain level of reliability (Southam-Gerow et al., 2020).

Therefore, this study aims to estimate the reliability of mathematical literacy assessment instruments for pre-service teachers using Generalizability Theory and to provide empirical recommendations for optimal assessment design through D-Study. The results are expected to contribute as a methodological guide for higher education institution in developing robust dan quality assured evaluation program to ensure the production of qualified educators.

METHODS

This study is part of an effort to develop a mathematical literacy assessment instrument for pre-service mathematics teachers. The purpose of this study is to evaluate the reliability of the scores obtained from this instrument, ensuring it possesses the adequate quality to measure the target competencies.

In this study, the reliability analysis was conducted using Generalizability Theory (G-Theory) and Dependability Study (D-Study). G-Theory developed by Cronbach et al, (1972), measures instrument reliability by simultaneously considering various sources of varians that affect assessment scores, while D-Study refers to the extent to which assessment results can be relied upon to make accurate and consistent decisions. Key considerations in using these theories include relative error variance and absolute error variance (Robert L. Brennan, 2001; Cronbach et al., 1972; Shavelson et al., 1991). Thus, it can be determined whether the assessment instrument used has adequate reliability to measure the mathematical literacy skills of pre-service mathematics teachers.

This study involved 30 undergraduate students (p) from the Mathematics Education Program at Universitas Pasir Pengaraian. Participans were in their second and third year of study and had completed foundational courses, ensuring they possessed the prerequisite knowledge to engage with the assessment.

The instrument was developed based on the framework of the International Association for the Evaluation of Educational Achievement (IEA). It consists 6 open-ended essay items (i) covering cognitive level of knowing, applying, and reasoning.

a. Knowing: recalling fact and performing procedures.
b. Applying: strategy selection and interpreting result.
c. Reasoning: analysis, conclusion drawing, and mathematical argumentation.

https://doi.org/10.30606/absis.v9i1.3585

In addition to referencing cognitive levels, the scope of mathematical content in the questions covers three main domains, namely (1) algebra and number operations; (2) geometry and measurement; (3) data and uncertainty. The questions were designed to be contextual and reflect the application of mathematics in everyday life. Due to the open-ended nature of the questions and the requirement for reasoning, the assessment process was conducted by several lecturer raters to increase objectivity and reduce bias in the assessment.

To address the subjective nature of essay scoring, the study employed a two-facet nested design $p \times (i{:}r)$. The implementation of G-theory in this study utilized a two-facet design, specifically an interaction between individuals (persons) and items nested within raters. The procedure followed these stages:

a. G-study: estimating the variance component associated with person, item, raters and their interaction.
b. D-study: using the variance component from the G-study to design an optimal measurement procedure by determining the necessary number of items and raters to achieve desired dependability level.

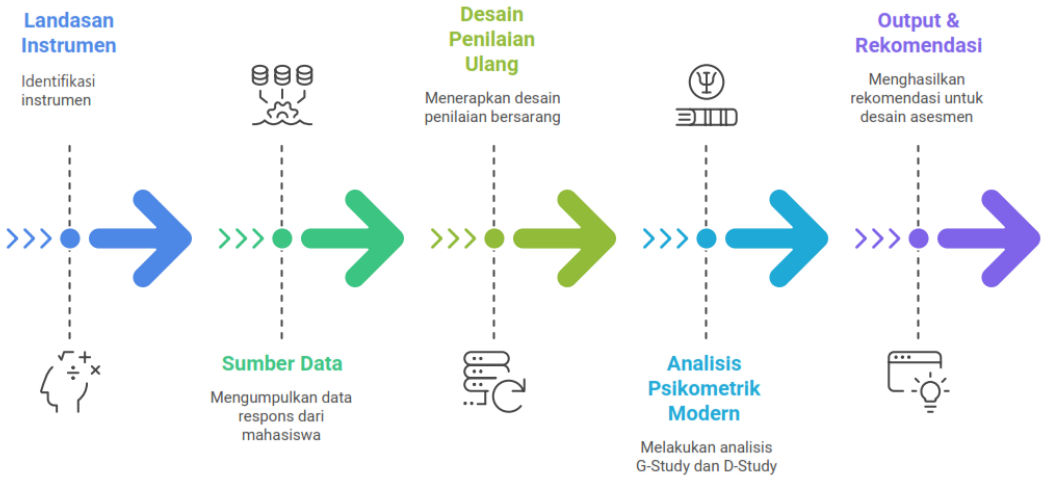

Figure 1. Research Procedure for Evaluating the Reliability of Instrumen

Data analysis was conducted using the "gstudy" and "dstudy" functions in the "gtheory" package in Rstudio (Huebner & Lucht, 2019; Moore, 2016).

FINDING AND DISCUSSION

Activities to determine the mathematical literacy skills of students are carried out through measurement activities. Any activity conducted in the world is inseparable from measurement (Mardapi, 2012). Measurement results can be trusted if repeated measurements on the same subject yield relatively similar results, which is known as reliability.

The procedure for measuring the mathematical literacy skills of pre-service mathematics teachers is influenced by several factors, including raters, items, and

their interactions. The mathematical literacy assessment instrument is a development of cognitive level items that encompass understanding and knowledge (knowing), application (applying), and reason (reasoning). Each cognitive level is evaluated from two assessment aspects and rated by a single rater. The items for each cognitive level are nested within raters, as described in the model depicted in Figure 1.

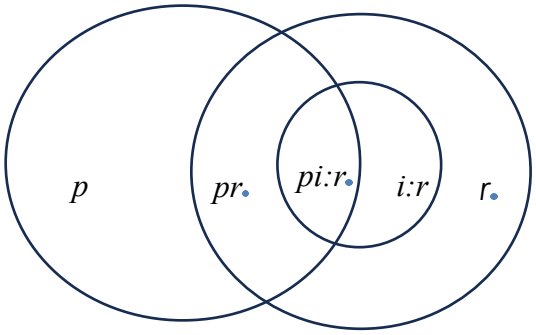

Figure 2. Mathematical Literacy Data Model Design (Robert L. Brennan, 2001)

Based on Figure 2, there are several sources of variance that can affect the measurement results, including person variance, rater variance, person-item interaction nested within raters, item-rater interaction, and residual variance. Addressing rater variability and consistency can help improve the reliability and generalizability of the measurement. The variance components for each source of variance are presented in Table 1 according to the model.

Table 1. ANOVA for p × (i:r) Mathematical Literacy Data

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	Variance	Percentage of Variance
Rater	2	2800.3	1400.2	23.02	92.8
Item: Rater	3	53.5	17.8	0.58	2.3
Student	29	65.36	2.254	0.10	0.4
StudentRater	58	96.02	1.656	0.56	2.2
StudentItem: Rater (Residual)	87	46.98	0.54	0.54	2.2

The rater has the highest variance of 23.02 with a percentage of variance of 92.8%. This indicates that the differences between raters are substantial and contribute significantly to the variation in scores, suggesting a high possibility of bias or inconsistency among raters in the scoring process. Meanwhile, the other variance components have very small percentages of variance. This suggests that the variance

https://doi.org/10.30606/absis.v9i1.3585

arising from students, student-rater interaction, and rater-item interaction is not significant.

The dominance of rater variance (92.8%) suggests a significant lack of inter-rater consistency, which can be attributed to the complexity of the open-ended mathematical literacy tasks. Since these items require reasoning and argumentation, raters may have applied different internal standards or weights to the students' qualitative explanations. This finding implies that the existing scoring rubric might be too general or open to multiple interpretations, failing to provide sufficiently granular criteria for each cognitive level, namely knowing, applying, and reasoning. Consequently, the high variance reflects "rater drift" or varying levels of stringency among the six lecturers involved.

The findings of this study offer crucial practical implications for the implementation of mathematical literacy assessment, particularly within teacher programs such as the one at Univeritas Pasir Pengaraian. The high rater variance underscores that in complex open-ended essay question, score consistency is highly susceptible to the individual subjectivity of the lecturer acting as evaluators. Practically, this indicated that institutions should bot rely on a single rater to evaluate pre-service teachers' if they aim to achieve high accuracy and fairness.

The results of the D-study are an important aspect in generalizability analysis. The D-study yielded a generalizability coefficient (G coefficient) of 0.27, indicating that the reliability of the measurement is relatively low. This means that the scores given by the raters do not fully reflect the students' actual mathematical literacy abilities due to the large variance caused by other factors. The dependability coefficient was found to be 0.01, which is very low. This suggests that the assessment of mathematical literacy is not dependable for making accurate decisions about students' mathematical literacy abilities.

The relative error variance of 0.28 and absolute error variance of 8.04 indicate that there is a considerable level of error in the measurement. The high Standard Error of Measurement (SEM) of 0.53 also suggests a large uncertainty in the scores obtained.

Some variance components of the assessment instrument are acceptable, but the large differences between raters and low reliability need to be improved to obtain more consistent and dependable assessment results. Table 2 presents several possibilities when the number of raters and items is increased or decreased.

Table 2. D Studi p×(i:r) for Mathematical Literacy Data

Source of Variation	$\hat{\sigma}^2$				
n rater	4	1	2	3	4
n item	6	4	6	4	6
Student	0.1	0.1	0.1	0.1	0.1
Rater	23.09	23.02	11.51	7.67	5.75

Source of Variation	$\hat{\sigma}^2$				
n rater	4	1	2	3	4
n item	6	4	6	4	6
Item:Rater	0.58	0.14	0.05	0.05	0.02
StudentRater	0.56	0.56	0.28	0.19	0.14
StudentItem:Rater	0.54	0.14	0.05	0.05	0.02
Relative Measurement Error Variance		0.69	0.32	0.23	0.16
G coefisien		0.13	0.24	0.3	0.38
Absolute Error Variance		23.86	11.88	7.95	5.94
Phi coefisien		0.00	0.01	0.01	0.02

Relative Error Variance and Generalizability Coefficient

The increase in the number of raters and items was accompanied by a decrease in relative error variance from 0.69 to 0.16. The significant decrease in relative error variance when the number of raters and items was increased indicates that adding raters and items is effective in reducing errors in scores relative to true variation.

Meanwhile, there was an increase in the G coefficient value from 0.13 to 0.38. However, the G coefficient value is still relatively low to moderate. A higher G coefficient value (close to 1) indicates that the measurement is more reliable and the results can be generalized better to various measurement conditions. In this condition, there is room to improve measurement reliability further by adding more raters or items.

Absolute Error Variance and Phi Coefficient

The addition of raters and items resulted in a significant decrease in absolute error variance from 23.86 to 5.94. This addition caused the absolute error in measurement to become smaller. This means that increasing the number of raters and items improves the absolute reliability of the measurement.

The phi coefficient value was very low, ranging from 0.00 to 0.02. This indicates that there are many sources of variation in the assessment that are not controlled. Thus, the assessment results from the raters cannot be relied upon absolutely to make important decisions.

The D-study results provide a strategic roadmap for department heads and quality assurance units. The data suggests that increasing the number of raters yields a more significant reduction in absolute error variance compared to merely increasing the number of test items. Therefore, a tangible implication for an Outcome Based Education (OBE) curriculum is the necessity of standardizing scoring procedures

https://doi.org/10.30606/absis.v9i1.3585

through the development of rigid rubrics and the implementation of regular rater training. Such measures are essential to align perceptions among evaluators before assessments take place, ensuring that academic decision regarding a candidate's qualification reflect their actual ability rather than a rater's level of leniency or bias.

CONCLUSIONS AND RECOMMENDATIONS

This study concludes that rater variability is the most dominant source of measurement error in the mathematical literacy assessment, accounting for 92.8% of the total score variance. Although the Decision Study (D-Study) results indicate that increasing the number of raters and items can theoretically reduce error components, the actual generalizability and dependability coefficients remain in a low category, ranging from 0.13 to 0.38. These findings provide critical evidence that the current assessment design is not yet stable or dependable enough for high-stakes decision-making or broad generalization. Instead, these findings serve as a vital diagnostic tool revealing that the assessment framework at Universitas Pasir Pengaraian requires fundamental refinement before it can be reliably implemented.

The implications for educational practice center on the urgent need to overhaul scoring protocols. The low level of dependability identified in this study implies that simply increasing the number of raters is insufficient if the underlying scoring rubrics remain subjective or poorly defined. This underscores the necessity of transitioning from traditional grading practices toward a more rigorous and standardized system. Such a shift includes the development of highly granular rubrics, specifically addressing cognitive levels of knowing, applying, and reasoning to minimize rater drift. Practically, these findings suggest that the current evaluation results should be used primarily for formative feedback rather than definitive competency certification until measurement stability is significantly improved.

Based on these findings, this study recommends a comprehensive revision of the mathematical literacy instrument and its accompanying rubrics. Immediate steps should include intensive rater training and calibration sessions to harmonize internal standards among lecturers. For future research, it is suggested to transition from a nested design to a fully crossed measurement design to better capture the interactions between items and raters. Furthermore, studies involving larger, multi-institutional samples are necessary to validate improved versions of the instrument. Future developments should also investigate the integration of structured scoring moderation processes to ensure that the evaluation of pre-service teachers meets the required standards of professional accountability.

REFERENCES

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Ayuningtyas, N., & Sukriyah. (2020). Analisis pengetahuan numerasi mahasiswa matematika calon guru. *Delta-Pi: Jurnal Matematika dan Pendidikan Matematika, 12*(1), 39–48.

Basri, H., Kurnadi, B., & Tafriliyanto, C. F. (2021). Investigasi kemampuan numerasi. *Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika, 4*(2), 72–79.

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Bulková, Z., Gašparík, J., Mašek, J., & Zitrický, V. (2022). Analytical procedures for the evaluation of infrastructural measures for increasing the capacity of railway lines. *Sustainability, 14*(21), Article 14430. https://doi.org/10.3390/su142114430

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons.

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research & Evaluation, 24*(5), 1–15.

Mardapi, D. (2012). *Pengukuran, penilaian, dan evaluasi pendidikan*. Nuha Medika.

Moore, C. T. (2016). *gtheory: Apply generalizability theory with R* [Computer software]. http://evaluationdashboard.com

Mullis, I. V. S., & Martin, M. O. (2017). *TIMSS 2019 assessment framework*. International Association for the Evaluation of Educational Achievement.

OECD. (2017). *PISA 2015 assessment and analytical framework: Science, reading, mathematics, financial literacy and collaborative problem solving* (Rev. ed.). OECD Publishing.

Ojose, B. (2011). Mathematics literacy: Are we able to put the mathematics we learn into everyday use? *Journal of Mathematics Education, 4*(1), 89–100.

Southam-Gerow, M. A., Bonifay, W., McLeod, B. D., Cox, J. R., Violante, S., Kendall, P. C., & Weisz, J. R. (2020). Generalizability and decision studies of a treatment adherence instrument. *Assessment, 27*(2), 321–333. https://doi.org/10.1177/1073191118765365

Tan, H. (2023). The generalizability of explanations. *Proceedings of the International Joint Conference on Neural Networks, 2023-June*. https://doi.org/10.1109/IJCNN54540.2023.10191972

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Practical Assessment, Research & Evaluation, 23*(1), 1–26.

Whitney-Smith, R., Hurrell, D., & Day, L. (2022). The role of mathematics education in developing students' 21st century skills, competencies and STEM capabilities. In *Proceedings of the annual conference of the Mathematics Education Research Group of Australasia* (pp. 554–561). Mathematics Education Research Group of Australasia. https://merga.net.au/Public/Public/Publications/Conference_Proceedings.aspx

Wijaya, A., & Dewayani, S. (2021). *Framework asesmen kompetensi minimum (AKM)*. Badan Penelitian dan Pengembangan dan Perbukuan, Kementerian Pendidikan dan Kebudayaan.