Psychometric Quality of the Non-Geometry Problem-Solving Instrument Adapted from Indonesia's MCA in Mathematics Literacy Using the Rasch Model

Fakhrur Razi[1]

[1]Faculty of Education, Universitas Islam Internasional Indonesia, Indonesia

Corresponding author: fakhrur.razi@uiii.ac.id

Abstract

This study developed a non-geometry problem-solving instrument by adapting mathematical literacy items from the Minimum Competency Assessment (MCA) Indonesia to address the limited empirical evidence of its psychometric quality. The development of the non-geometry problem-solving instrument begins with a theoretical framework, item collection, review process, and pilot testing. The researcher administered the 30-item instrument to 78 students in 8th-grade secondary schools via convenience sampling, focusing on evaluating reliability, item fit, and unidimensionality. The analysis showed that the Cronbach Alpha score is 0.66 (modest), the Person Reliability score is 0.54 (low), and the Item Reliability score is 0.90 (high). Based on item fit, one item does not fit the model because it is extremely difficult, so the respondents tend to guess the answer. Meanwhile, the Rasch model has explained 19.8% (approximately 20%) of the total variance, indicating insufficient unidimensionality due to the complexity of the competencies involved and the small sample size used to respond to the items. In conclusion, most items have good qualities according to the Rasch model, although some need further evaluation. This finding also suggests the importance of testing the instrument on a wider scale to enhance psychometric quality.

Keywords: Psychometric Quality, Non-Geometry Problem Solving Instrument, Indonesia's MCA, Rasch Model

Introduction

The Minimum Competency Assessment (MCA) in mathematics literacy is part of the national assessment, a standardised test administered at all school levels in Indonesia since 2021 (MoEC, 2021). Initially, the subject-based national examination was used for years to monitor the development of education in Indonesia (Berkhout, 2014). However, the government has considered the competency assessment test, such as the PISA (Programme for International Student Assessment), to be more relevant in testing students basic competencies (MoEC , 2020). This test emerged due to concerns over Indonesia's stagnating performance in international benchmarks. For example, Indonesia scored 366 in mathematics, significantly lower than the OECD countries average of 489 in 2018 (OECD, 2019), signalling fundamental challenges in students' mathematics literacy.

Despite this conceptual alignment, the MCA differs substantially from established international assessments in terms of psychometric quality and data transparency. International assessments, such as TIMSS (Trends in International Mathematics and Science Study) and PISA, undergo rigorous validation processes to ensure they accurately measure students competencies across diverse educational contexts (Wu, 2010). Furthermore, the results of Indonesia 's MCA are not publicly disseminated but are instead reported to schools as aggregate averages (MoEC, 2020). This lack of transparency highlights a critical gap in research regarding the quality of MCA items, raising questions about their effectiveness in assessing students mathematical literacy.

In this study, a non-geometry problem-solving instrument will be developed to address broader research and assessment needs. The geometry domain was intentionally excluded from the instrument because it was not required for subsequent research. In addition, adapting items derived from a standardised test is crucial because the test has undergone development (item development, internal consistency, validity) and widespread use (Crocker & Algina, 2006). These needs have been met by established instruments, such as PISA and TIMSS, which are used to measure the quality of the education system in several countries (OECD, 2019; Philpot et al., 2021).

In this development process, the quality of the items adapted from Indonesia's MCA for mathematics literacy needed to be tested, as the instrument has not been sufficiently psychometrically validated. Therefore, this study aimed to answer the research question: *To what extent do the Non-Geometry Problem-Solving Instrument items demonstrate acceptable psychometric quality when adapted from Indonesia's MCA in Mathematics Literacy?* Through this development and evaluation using the Rasch Model, the researcher obtained evidence that the instrument can measure non-geometry problem-solving competency among 8th-grade secondary students in the Indonesian context.

Conceptual Framework

Mathematics Literacy

Mathematical literacy is the capacity to identify and understand the role of mathematics in everyday life, including awareness of its role, decision-making, and problem-solving (OECD, 2019). In Indonesias MCA framework, mathematical literacy comprises several stages that follow the PISA framework (OECD, 2019), namely formulating mathematical situations, employing mathematical concepts, and interpreting, applying, and evaluating mathematical outcomes. These processes are interconnected to solve contextual mathematical problems. In addition, the framework also uses cognitive level adopting the TIMSS framework (Philpot et al., 2021), namely knowing, formulating, and reasoning. All three include the thinking processes students need to understand, apply, and reason about mathematical concepts. In this study, there are 3 mathematical domains taken (excluding the geometry domain), explained in detail in Figure 1.

Figure 1

Indicators of Indonesia's MCA in mathematics literacy (Grade 8)

Domain	Indicators
Numbers	1. Using addition/subtraction/multiplication/division of fractions or decimal numbers, including calculating the square and cube of a decimal number with one digit after the decimal point. As well as operations on integers, including negative integers (including estimating the results of operations). 2. Determining the Lowest Common Multiple (LCM) and Highest Common Factor (HCF). 3. Understanding integers, powers, and roots. 4. Sorting numbers including negative integers, decimals, percentages, and fractions. 5. Generalizing patterns of number sequences and object configurations.
Algebra	1. Determining the nth term in a simple number pattern. 2. Solving linear equations and inequalities in 1 variable and systems of linear equations in 2 variables. 3. Solving social arithmetic problems involving ratios/percentages. 4. Solving linear inequalities in 1 variable or systems of linear equations in 2 variables. 5. Recognize and continue patterns of increasing and decreasing numbers involving multiplication and division.
Data and Uncertainty	1. Determining and using mean, median, and mode in problem solving. 2. Calculating the probability of simple events. 3. Reading (= extracting information from) data presented in the form of tables, bar charts, line charts, and pie charts (including how to collect data and how to present it). 4. Presenting, analysing, and interpreting data in the form of bar charts or tables.

Note. Adapted from Pusat Asesmen Pendidikan (http://pusmendik.kemdikbud.go.id).

Item Development

The item development stage includes designing a theoretical framework, collecting items, reviewing items, and conducting pilot testing (Irwing & Hughes, 2017; Haladyna & Rodriguez, 2013). In this study, the researcher adapts theoretical frameworks for developing test items to ensure that the items align with the targeted competencies and cognitive demands relevant to non-geometry problem solving. These frameworks provide the foundation for defining the constructs to be measured and for guiding item construction to ensure construct validity (Messick, 1995).

Based on these frameworks, items are collected and designed to accurately reflect the constructs being measured. The quality of these items is then examined through a structured item review process involving expert judgment. This step is crucial for evaluating content quality, clarity, and the absence of ambiguity (NRC, 2004). Following the review, pilot testing is intended to obtain data to evaluate the quality of the items. This stage enables the collection of empirical data, which is essential for analysing the psychometric properties of the items (Germain, 2006).

Item Quality

In this study, psychometric quality, as assessed by the Rash Model, is determined by reliability, item fit, and unidimensionality. Reliability in the Rasch context refers to the internal consistency and replicability of the measurement, from the perspectives of both persons (respondents) and items (Bond & Fox, 2015). Specifically, person reliability reflects the degree to which the instrument can distinguish between high- and low-ability respondents. In contrast, item reliability refers to the precision with which item difficulties are estimated (Bond & Fox, 2015). High reliability indicates that the instrument yields stable, consistent measurements across administrations and samples (Wright & Masters, 1982).

Item fit refers to the extent to which the observed response patterns align with the expectations of the Rasch Model (Bambang & Widhiarso, 2015). Items that exhibit misfit may not conform to the underlying measurement construct and could threaten the scales validity. Fit statistics, such as infit and outfit mean square values, are used to assess whether an item behaves predictably with respect to the latent trait being measured (Boone et al., 2014; Linacre, 2002). Items with acceptable fit statistics contribute meaningfully to the measurement, whereas items with poor fit may introduce noise or multidimensionality.

Unidimensionality is another critical assumption of Rasch measurement. It implies that all items on the scale measure a single underlying construct (Wind & Hua, 2021). This is essential to ensure that the total test score represents one coherent domain rather than multiple unrelated traits (Wright & Stone, 1999; Tennant & Pallant, 2006). Unidimensionality is typically assessed through principal component analysis (PCA) of residuals. If the residuals do not form meaningful secondary dimensions, it supports the claim that the instrument is unidimensional.

Methodology

Item Development and Sample Selection

The instrument development process comprised several stages, as explained in the item development section above. Then, the quality of this instrument was tested using data from 78 Grade 8 students selected through convenience sampling from two public schools in Depok City, which were randomly selected. Then, one available class was selected from each school, and all students from that class were included as participants.

Participants were drawn from two randomly selected public secondary schools in Depok City, West Java, Indonesia. From each selected school, one intact class available at the time of data collection was chosen, and all students in that class were included as respondents. This class-based sampling approach is commonly employed in educational settings (George, 2021) because it allows researchers to preserve students natural groupings and minimise disruptions to instructional activities.

Evaluation of Items

The Rasch model, using the WINSTEPS Version 3.73 application, will be used to evaluate item quality. The researcher will evaluate several aspects related to the instruments reliability, item fit, and unidimensionality. This analysis is used to ensure that the items have psychometric properties consistent with the Rasch model and to evaluate item improvements. Reliability was assessed at both the person and item levels, with higher indices indicating that the instrument can consistently differentiate between varying levels of student ability. That item difficulty estimates are stable and reproducible (Wright & Masters, 1982).

Item fit was evaluated using infit and outfit mean square (MNSQ) statistics, which assess how well each item aligns with the Rasch models expectations (Linacre, 2002). Misfitting items can distort measurement and signal multidimensionality, which compromises construct validity (Boone et al., 2014). To confirm unidimensionality, a core assumption of Rasch analysis, the study used Principal Components Analysis (PCA) of residuals. This technique tests whether all items in the scale measure a single latent trait, which is essential for meaningful Rasch-based interpretation (Tennant & Pallant, 2006).

Key Findings and Discussion

Description of Items

The researcher developed a non-geometry problem-solving instrument by collecting items from the official website of Indonesias Educational Assessment Centre. These items were aligned with the PISA mathematics framework, distributing the questions across three problem-solving processes: formulating situations mathematically, employing mathematical concepts and reasoning, and interpreting and evaluating outcomes, which reflects the typical distribution in PISA (OECD, 2017; OECD, 2022). Additionally, the items were analysed using the TIMSS cognitive framework, which classifies them into the domains of knowing, applying, and reasoning based on item indicators (Philpot et al., 2021; Kemendikbud, 2021). This dual-framework approach ensures that the instrument comprehensively captures students mathematical problem -solving skills in non-geometry contexts.

Each item was selected and revised, resulting in a total of 30 items across the domains of number, algebra, data, and uncertainty. The item formats varied and included multiple-choice questions with single or multiple answers, short-answer responses, and matching tasks, all of which incorporated real-world contexts with data presented in figures, tables, or diagrams. The researcher analysed the distribution of problem-solving processes across the items based on their underlying intent and categorised each item according to cognitive levels using established indicators. The items were subsequently reviewed by several experts to evaluate language clarity, logical consistency, and content relevance. The distribution of the 30 items according to the framework is presented in Table 1.

Table 1

Proportion of Non-Geometry Problem-Solving Questions Based on the Framework

Criteria	Level	Percentage	Number of Items
Mathematical Problem-Solving Process (Philpot et al., 2021; MoEC, 2021)			
	Formulating	25%	8
Question intent of each item	Employing	50%	15
	Interpreting	25%	7
Cognitive Process (OECD, 2019; MoEC, 2021)			
	Knowing	30%	9
Indicators of each item	Applying	55%	17
	Reasoning	15%	5

The Quality of Items

The researcher administered the instrument to 78 students, using 30 revised items from the non-geometry problem-solving test. Below is the data analysis of the non-geometry problem-solving test instrument, which includes reliability, item fit, and unidimensionality.

a) Reliability

This statistic provides an overview of how the items perform under the Rasch model, including the instruments reliability. In this study, reliability is the primary component of the instruments quality, supported by other statistical measures, as detailed in Figure 2.

Figure 2

Summary Statistics of the Instrument

SUMMARY OF 78 MEASURED Person

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	9.7	29.2	-.91	.46	1.00	.0	1.01	.0
S.D.	4.0	3.3	.71	.10	.16	.9	.36	.9
MAX.	20.0	30.0	.81	1.06	1.39	1.8	2.51	2.4
MIN.	1.0	15.0	-2.81	.40	.68	-2.4	.44	-2.2

REAL RMSE .48 TRUE SD .52 SEPARATION 1.09 Person RELIABILITY .54
MODEL RMSE .47 TRUE SD .54 SEPARATION 1.15 Person RELIABILITY .57
S.E. OF Person MEAN = .08

VALID RESPONSES: 97.4% (APPROXIMATE)
Person RAW SCORE-TO-MEASURE CORRELATION = .98 (approximate due to missing data)
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .66 (approximate due to missing data)

SUMMARY OF 30 MEASURED Item

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	25.2	76.0	.00	.28	1.00	.0	1.01	-.1
S.D.	12.3	2.0	.92	.06	.11	1.1	.21	1.1
MAX.	52.0	78.0	1.95	.47	1.18	2.5	1.75	2.5
MIN.	5.0	73.0	-1.68	.24	.85	-1.8	.62	-1.7

REAL RMSE .30 TRUE SD .87 SEPARATION 2.94 Item RELIABILITY .90
MODEL RMSE .29 TRUE SD .87 SEPARATION 3.02 Item RELIABILITY .90
S.E. OF Item MEAN = .17

In Figure 2, the Cronbachs Alpha score is 0.66, which is above the commonly accepted minimum value of 0.60 for internal consistency (Bond & Fox, 2015), indicating a modest level of reliability and sufficient interaction between persons and test items. This indicates that the average of all the items on the instrument can measure non-geometry problem-solving as a construct. The alpha reliability falls in the modest range because the construct is too broad and the items are not clearly defined.

However, the Person Reliability score = 0.54 (<0.6, indicating low) is due to students abilities which are still below the level of difficulty of the items (the Person Measure = -0.91 logit <0.0 logit) so that students abilities are less spread out (the Person Separation score = 1.09 <1.5), and only distribute in 1 or 2 groups of student's ability (Boone et al., 2014). In contrast, an Item Reliability score of 0.9 indicates a high level of consistency among items across students (Bambang & Widhiarso, 2015). This is also relevant to the Item Separation score of 2.94 (>2), indicating that item difficulty is well distributed and can be categorised into 4 groups (Boone et al., 2014).

b) Item Fit

The following data from item fit in Figure 3 show the suitability of the items with the Rasch model. According to Bond and Fox (2015), item fit analysis helps identify inappropriate or problematic items that can affect the instruments overall validity and reliability. The data is explained as follows:

Figure 3

Item Fit of the Instrument

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
5	8	78	1.44	.38	1.11	.5	1.75	1.7	A-.07	.20	89.7	89.7	s5
27	22	74	.08	.27	1.16	1.4	1.29	1.8	B .02	.29	68.9	71.5	s27
12	15	78	.66	.30	1.15	.9	1.27	1.1	C .02	.26	82.1	80.8	s12
13	12	78	.95	.32	1.14	.7	1.24	.9	D .03	.24	84.6	84.6	s13
18	38	74	-.92	.24	1.18	2.5	1.22	2.5	E .04	.31	51.4	62.7	s18
7	26	78	-.15	.25	1.13	1.4	1.19	1.4	F .11	.30	65.4	68.9	s7
26	5	74	1.95	.47	.89	-.2	1.19	.5	G .28	.17	93.2	93.2	s26
14	17	78	.49	.29	1.15	1.0	1.11	.6	H .09	.27	75.6	78.4	s14
15	26	78	-.15	.25	1.12	1.2	1.10	.8	I .15	.30	60.3	68.9	s15
22	39	74	-.98	.24	1.10	1.4	1.09	1.1	J .17	.31	55.4	62.8	s22
30	15	74	.63	.30	1.09	.6	1.03	.2	K .15	.26	77.0	79.8	s30
11	16	78	.58	.29	1.05	.4	1.06	.3	L .19	.26	79.5	79.6	s11
3	8	78	1.44	.38	1.03	.2	.95	.0	M .17	.20	89.7	89.7	s3
24	46	74	-1.41	.25	1.03	.4	1.00	.0	N .27	.30	60.8	66.3	s24
28	43	74	-1.22	.25	1.01	.2	.99	-.1	O .30	.31	60.8	64.3	s28
6	35	78	-.68	.24	1.00	.1	1.01	.1	o .31	.32	70.5	63.6	s6
29	24	74	-.05	.26	1.00	.0	.93	-.5	n .32	.29	66.2	69.3	s29
10	18	78	.41	.28	.97	-.1	.92	-.4	m .32	.27	78.2	77.3	s10
19	34	74	-.68	.25	.97	-.4	.96	-.4	l .35	.31	67.6	62.9	s19
25	28	74	-.32	.25	.94	-.6	.90	-1.0	k .40	.30	67.6	66.1	s25
1	52	78	-1.68	.25	.94	-.6	.90	-.8	j .40	.31	70.5	69.4	s1
20	26	74	-.19	.26	.89	-1.2	.85	-1.3	i .46	.30	73.0	67.5	s20
17	40	74	-1.04	.25	.88	-1.7	.87	-1.6	h .47	.31	75.7	63.1	s17
21	25	74	-.12	.26	.88	-1.3	.81	-1.6	g .49	.29	70.3	68.4	s21
4	42	78	-1.09	.24	.87	-1.8	.87	-1.5	f .49	.32	74.4	63.6	s4
2	25	78	-.08	.25	.87	-1.3	.84	-1.2	e .47	.30	76.9	69.8	s2
9	20	78	.26	.27	.86	-1.0	.77	-1.3	d .48	.28	76.9	75.0	s9
16	6	73	1.75	.43	.86	-.3	.62	-.8	c .41	.18	91.8	91.8	s16
23	18	74	.38	.28	.85	-1.1	.85	-.8	b .47	.27	79.7	76.1	s23
8	28	78	-.27	.25	.85	-1.7	.80	-1.7	a .52	.31	79.5	67.4	s8
MEAN	25.2	76.0	.00	.28	1.00	.0	1.01	-.1			73.8	73.1	
S.D.	12.3	2.0	.92	.06	.11	1.1	.21	1.1			10.3	9.3	

Based on the item fit table, at least two criteria below have to be met, taking into account several criteria used to determine an item fit (Sumintono & Widhiarso, 2015):

1) 0.5 < OUTFIT MNSQ < 1.5 (Bond & Fox, 2015),
2) -2 < OUTFIT ZSTD < 2 Wright & Linacre (1994),
3) 0.4 < Pt Measure Corr < 0.85 (Bond & Fox, 2015).

Based on the criteria, the Rasch analysis indicates that students responses to some items are inconsistent with the model. Several factors that cause this to happen can occur either due to student factors (such as guessing and carelessness) or to the item itself (multi-interpretation, ambiguity, and unclear wording). In this instrument, item number 5 (s5) is outside the acceptable range, with the OUTFIT MNSQ score = 1.75 and the Pt Measure Corr. score = - 0.7. The item fit data indicate that this item is too difficult for the students, prompting them to guess the answer (Wright & Linacre, 1994).

c) Unidimensionality

The assumption of unidimensionality in the Rasch model ensures that all items measure the same construct (Kline, 2005). To evaluate this assumption, a principal component analysis was performed on the residuals as follows:

Figure 4

Standardised Residual Variance of the Instrument

```
INPUT: 78 Person  30 Item  REPORTED: 78 Person  30 Item  2 CATS    WINSTEPS 3.73
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                                -- Empirical --    Modeled
Total raw variance in observations     =        37.4 100.0%         100.0%
  Raw variance explained by measures   =         7.4  19.8%          19.7%
    Raw variance explained by persons  =         1.5   4.0%           3.9%
    Raw Variance explained by items    =         5.9  15.8%          15.7%
  Raw unexplained variance (total)     =        30.0  80.2% 100.0%   80.3%
    Unexplned variance in 1st contrast =         3.1   8.3%  10.4%
    Unexplned variance in 2nd contrast =         2.5   6.8%   8.5%
    Unexplned variance in 3rd contrast =         2.2   5.9%   7.3%
    Unexplned variance in 4th contrast =         2.0   5.4%   6.7%
    Unexplned variance in 5th contrast =         1.9   5.0%   6.2%
```

Figure 4 shows that the Rasch model explains 19.8% (approximately 20%) of the total variance, which approaches sufficiency, particularly when the constructs measured are complex, and the sample size is limited (Linacre, 2011; Bond & Fox, 2015). Mathematical literacy test items assess complex, interrelated components because students are expected to integrate various knowledge and skills to solve problems (OECD, 2019). For example, in this instrument, the ability to read data in tables or diagrams is necessary to understand the context fully. Measurement errors can also indicate unexplained variance, as shown by the 80.2% of the variance not explained by the model, attributable to incorrect answers (Linacre, 2011).

The first contrast eigenvalue of 3.1 far exceeds the commonly used threshold of >2.0, indicating the potential for additional dimensions to emerge beyond the primary construct (Smith et al., 2008). Therefore, further evaluation of the items contributing to the contrast value is needed. Furthermore, the items influencing the second and third contrasts still need to be evaluated because their values slightly exceed the threshold. These findings still suggest the presence of additional, potentially meaningful residual patterns. However, the primary dimensions explain enough variance to indicate that this instrument measures a single construct. However, the items need to be refined so that the measurement is more focused on one ability constructed (Wagner-Menghin, 2007; Bond & Fox, 2015).

Conclusion

The Rasch analysis indicates that the developed non-geometry problem-solving items had moderate reliability and well-distributed item difficulty levels. However, the instrument was still limited in differentiating student abilities due to the high difficulty of the items and the small sample size. Although most items fit the model, one item did not, and the high eigenvalue of the first contrast indicated the need for item refinement to improve unidimensionality and measurement precision. Thus, these findings suggest the need to evaluate problematic items and conduct large-scale pilot testing to develop a more robust instrument.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The author(s) acknowledge the use of ChatGPT (OpenAI) in the preparation of this writing process. ChatGPT was used to support the manuscript by assisting with drafting, paraphrasing, translating, clarifying language, and improving the texts readability.

References

Berkhout, E., Pradhan, M., Suryadarma, D., & Swarnata, A. (2024). Using technology to prevent fraud in high-stakes national school examinations: Evidence from Indonesia. *Journal of Development Economics, 170,* 103307. https://doi.org/10.1016/j.jdeveco.2024.103307

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences.* Springer.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Cengage Learning.

Germain, M. L. (2006). *Stages of Psychometric Measure Development: The Example of the Generalised Expertise Measure (GEM).* Online Submission.

George, A. C. (2021). A brief guide to sampling in educational settings. *The Quantitative Methods for Psychology*, *14*(3), 286–298.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hathcoat, J. D., Sanders, C. B., & Gregg, N. (2016). *Selecting and designing instruments: Item development, reliability, and validity.* Centre for Assessment and Research Studies, James Madison University.

Irwing, P., & Hughes, D. J. (2017). *Test Development.* University of Manchester Research.

Kline, P. (2005). *Psychological testing: A practical approach to design and evaluation.* Sage Publications.

Linacre, J. M. (2011). *A user's guide to Winsteps: Rasch-model computer programs.* Winsteps.com.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as a scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Ministry of Education and Culture (MoEC). (2020). (2021). *Frequently asked questions about the national assessment* [Lembar tanya jawab asesmen nasional]. Centre for Assessment and Learning, Agency for Standards, Curriculum, and Educational Assessment.

Ministry of Education and Culture (MoEC). (2021). *Minimum Competency Assessment Framework (AKM)* [Framework Asesmen Kompetensi Minimum]. Centre for Assessment and Learning, Agency for Research and Development and Books.

National Research Council (NRC). (2004). *Redesigning the U.S. naturalisation tests: Interim report.* The National Academies Press. https://doi.org/10.17226/11168

OECD. (2019). *PISA 2018 assessment and analytical framework: Mathematics, reading, science and global competence.* OECD Publishing. https://doi.org/10.1787/b25efab8-en

Philpot, R., Lindquist, M., Mullis, I. V., & Aldrich, C. E. (2021). TIMSS 2023 Mathematics Framework. *TIMSS 2023 Assessment Frameworks, 5.*

Smith, R. M., Schumacker, R. E., & Bush, M. J. (2008). Using item mean squares to evaluate fit to the Rasch model. *Journal of Applied Measurement, 9*(4), 318–334.

Sumintono, B., & Widhiarso, W. (2015). *Application of the Rasch model in educational assessment* [Aplikasi pemodelan Rasch pada assessment pendidikan]. Cimahi: Trim Komunikata Publisher.

Wagner-Menghin, M. M. (2007). Conception and construction of a Rasch-scaled measure for self-confidence in ones vocabulary ability. *Journal of Applied Measurement*, *8*(1), 35–47.

Wind, S., & Hua, C. (2021). *Rasch measurement theory analysis in R: Illustrations and practical guidance for researchers and practitioners*. Bookdown. org [Epub]

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370–371.

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2nd ed.). Wide Range Inc.

Wu, M. (2010). *Comparing the similarities and differences of PISA 2003 and TIMSS* (OECD Education Working Papers No. 32). OECD Publishing. https://doi.org/10.1787/5km4psnm13nx-en

Acknowledgment

Fakhrur Razi, Faculty of Education, Universitas Islam Internasional Indonesia, Indonesia
Email: fakhrur.razi@uiii.ac.id