

Construct validity and reliability of a high-stakes mathematics test for junior high school: A case from the Indonesian national examination

Astutiati¹; Muhammad Kusnun Muhsin¹; Endang Mulyatiningsih^{1*}; Heri Retnawati¹; Ezi Apino¹; Ibnu Rafi²; Umi Farisiyah¹

¹Universitas Negeri Yogyakarta, Indonesia

²National Sun Yat-sen University, Taiwan, Province of China

*Corresponding author. E-mail: rosnawati@uny.ac.id

ARTICLE INFO

Article History

Submitted:

10 February 2026

Revised:

26 June 2026

Accepted:

15 July 2026

Keywords

confirmatory factor analysis; construct validity; high-stake mathematics test

Scan Me:



ABSTRACT

High-stakes tests are used to inform critical decisions concerning students, teachers, schools, and educational policies. Therefore, they must meet rigorous quality-assurance standards. One key approach to ensuring the quality of high-stakes tests is establishing construct validity and reliability. This study aims to prove the construct validity and estimate the construct reliability for a high-stakes mathematics test for junior high school students. This study uses response data from participants taking the national junior high school mathematics exam in Indonesia. A simple random sampling technique was used to select 300 participants. Data were analyzed using second-order confirmatory factor analysis with the help of the R program. The study results revealed four factors that make up the mathematical ability construct: number, algebra, geometry, and statistics. All manifest variables (18 indicators) confirmed a significant contribution. The model fit indicators are met ($\chi^2 = 140.013$, $df = 131$, $p = 0.279$, $RMSEA = 0.015$, $CFI = 0.994$, $TLI = 0.993$; $SRMR = 0.039$), so the construct measured in the high-stakes mathematics test is valid for measuring students' mathematical abilities. Thus, the test construct has a reliability coefficient of 0.98, reflecting that the construct is reliable for measuring mathematical abilities.

This is an open access article under the [CC-BY-SA](#) license.



To cite this article (in APA style):

Astutiati, A., Muhsin, M. K., Mulyatiningsih, E., Retnawati, H., Apino, E., Rafi, I., & Farisiyah, U. (2026). Construct validity and reliability of a high-stakes mathematics test for junior high school: A case from the Indonesian national examination. *Jurnal Penelitian dan Evaluasi Pendidikan*, 30(1), 1–12. <https://doi.org/10.21831/jpep.v30i1.95822>

INTRODUCTION

The quality of educational programs can be assessed through the implementation of high-stakes testing policies. Therefore, high-stakes assessments are used to make significant educational decisions about students, teachers, schools, or school districts (Demir & Keleş, 2021; Gregory & Clarke, 2003; Heubert, 2000; Minarechová, 2012; Stobart & Eggen, 2012). This indicates that these tests influence school systems, teachers, students, politicians, and parents, both positively and negatively (Minarechová, 2012). Furthermore, high-stakes testing is linked to concepts such as school accountability, funding, and parental school choice (Minarechová, 2012; Nichols et al., 2012; Stobart & Eggen, 2012). One of the goals of high-stakes testing is to improve the performance of educational institutions. In addition, high-stakes testing is seen as a means to raise academic standards, hold educators and students accountable for meeting those standards, and increase public trust in schools (Heubert & Hauser, 1999; Minarechová, 2012; Nichols et al., 2012).

High-stakes testing is the process of attaching significant consequences to standardized test performance with the goal of boosting teacher effectiveness and student achievement, as [Nichols et al. \(2012\)](#) explored. However, the rationale is that by attaching significant rewards or serious threats to changes in student test scores, teachers and their students are likely to be motivated to work harder, improve, and learn more ([Nichols et al., 2012](#)). In practice, high-stakes testing typically involves national or regional standardized achievement tests ([Marchant, 2004](#)). High-stakes testing can consist of a single measure, such as a paper-and-pencil test, or multiple measures, including portfolios, projects, and written test scores ([Gregory & Clarke, 2003](#)). This testing practice is complex, characterized by the intersection of various interpretations, uses, stakeholders, and contexts ([Koch & DeLuca, 2012](#)). With the advancement of technology, this test is commonly administered through computer-based exams ([Retnawati, Hadi, et al., 2017](#)). High-stakes testing has become part of the education system in many countries ([Demir & Keleş, 2021](#); [Koch & DeLuca, 2012](#); [Minarechová, 2012](#)). Several developed and developing countries, such as England, the United States of America, Australia, Sweden, Chile, and Turkey, including Indonesia, have utilized this test ([Demir & Keleş, 2021](#)).

In Indonesia, high-stakes testing uses a national exam format for junior high and senior high school levels. At the junior high level, the national exam assesses student achievement in four subjects: Indonesian, English, mathematics, and science. At the high school level, the exam tests students' achievement in Indonesian, English, and major subjects (science, social studies, or language). This study focuses on junior high school students and mathematics. One important issue related to high-stakes testing is validity ([Koch & DeLuca, 2012](#); [Lane, 2004](#); [Smith & Fey, 2000](#)). Given the evolving nature of high-stakes testing and its impact on teaching and learning, the validation process is crucial to ensure tests are interpreted and used appropriately ([Koch & DeLuca, 2012](#); [Moss, 2007](#); [Stobart, 2009](#)). In addition to validity, [Stobart and Eggen \(2012\)](#) note that high-stakes test reliability is crucial. In assessment, validity and reliability are essential to ensure the quality of test instruments, including high-stakes tests.

Furthermore, in Indonesia, issues related to the validity and reliability of high-stakes tests are still rarely discussed. Several studies on the topic of national exams in Indonesia focus more on the challenges and changes in the implementation mode of national exams. For example, [Retnawati, Kartowagiran, et al. \(2017\)](#) investigated the causes of students' difficulties in answering items on the national mathematics exam and potential strategies that teachers can implement to overcome them. [Retnawati, Hadi, et al. \(2017\)](#) conducted research to uncover the challenges and strategies for implementing computer-based national exams in Indonesia. Furthermore, [Kartowagiran et al. \(2018\)](#) investigated the quality aspects of the tests used in national exams in Indonesia, namely whether the test kits consisting of several question packages are equivalent. [Retnawati et al. \(2019\)](#) investigated the impact of national exams no longer being the determinant of student graduation. The impact examined in this study focused on students' seriousness in preparing for and when they took the test. Still related to difficult items in the national mathematics exam, [Retnawati et al. \(2021\)](#) mapped mathematics topics considered difficult by national exam participants. More recently, a study by [Rafi et al. \(2023\)](#) aimed to uncover the characteristics of tests and items used in national exams in Indonesia. Although this study attempted to uncover aspects of the quality of national exam tests, its exploration did not focus on proving the construct validity and reliability of the tests.

Previous literature on the topic of national exams appears to have neglected to address the validity and reliability of tests used in national exams. Therefore, this study attempts to fill this gap. In addition, this study also provides guidance for test developers in developing standardized tests, particularly for mathematics, and serves as a reference for educators in designing mathematics learning according to expected core competencies. Based on this gap study, the purpose of this study is to demonstrate the construct validity and estimate the construct reliability of a high-stakes test (i.e., the national mathematics exam) for junior high school students.

RESEARCH METHOD

Design

This exploratory study aims to describe the quality of the standardized mathematics test used in the national junior high school exam in Indonesia. This test is considered high-stakes because its results have significant consequences for students, teachers, schools, and education stakeholders in Indonesia. The study used data from junior high school students' responses to the national mathematics exam from five provinces in Indonesia. The test consists of five different question sets, with the present study focusing solely on the first set. However, all five sets of questions were developed using the same blueprint. Therefore, all sets are considered equivalent. This study also aims to describe the quality of the high-stakes mathematics test based on two key aspects: evidence of construct validity and estimates of construct reliability. In line with these objectives, confirmatory factor analysis (CFA) is employed to examine whether the test construct is supported by empirical data. In addition, construct reliability is estimated to obtain information on the accuracy of the measurement results.

Participant

This study used responses from 8,480 students taking the national junior high school exam in Indonesia. This sample size is relatively large and could potentially result in a statistical model that does not fit the empirical data. Experts recommend using a smaller sample size to avoid the possibility of data not fitting the statistical model (Kyriazos, 2018). In CFA and SEM parameter estimation, the chi-square test and goodness of fit indices are both sensitive to sample size, so a smaller sample size is recommended (Kyriazos, 2018). Generally, CFA/SEM are large-sample techniques (Kline, 2016; Kyriazos, 2018), but as a rule, models with robust parameter estimates and variables with high reliability may require a smaller sample (Tabachnick & Fidell, 2013). DeVellis (2017) suggests that $N = 300$ is sufficient for a unidimensional scale consisting of a set of 20 items. Tabachnick and Fidell (2013) also suggested a sample size of 300 for CFA and SEM analysis. Based on this recommendation, Resampling was conducted to select 300 observations using a simple random sampling technique, ensuring that each student response had an equal probability of being included in the sample. The randomization procedure was carried out using the R program.

Instrument

This study used a standardized mathematics test used in the national junior high school exam in Indonesia. The primary construct measured in this test is junior high school students' mathematical abilities. The test measures four content aspects, namely number, algebra, geometry and measurement, and statistics and probability. Each aspect is broken down into item indicators. A total of 18 indicators are used to measure students' mathematical abilities. In this test, some indicators are measured by more than one item. This reflects the complexity and depth of the content measured by the indicators. The test has 40 multiple-choice items with four answer options and only one correct answer.

Research Procedure

This study utilized national mathematics exam results data for junior high school students. The data consisted of student responses after they took the test. Data access permission was granted by the Center for Research and Education (currently the Center for Educational Assessment), an institution authorized by the Ministry of Education, Culture, and Higher

Education to administer national examination data in Indonesia. Researchers were granted access to document student response data from five provinces in Indonesia. The data we received consisted of participant IDs, provincial codes (for the five provinces), and student responses to 40 test items. The student responses were dichotomized, with 1 representing a correct answer and 0 representing an incorrect answer. To protect student and school privacy, researchers were not granted access to the students' names and schools of origin. Researchers were only given access to the blueprint and one of the question packages (the main package) used in the national exam. Therefore, the blueprint and question package to determine the items for each indicator measured by the test.

Data Analysis

Descriptive analysis was used to determine the feasibility and characteristics of the data. Second, this study conducted factor analysis to prove the construct validity of the high-stakes mathematics test. In addition, second-order CFA was used to test the hypotheses. Through second-order CFA, we evaluated the goodness of fit of the conceptual model and factor loadings for each manifest variable. The goodness of fit was used to evaluate how well the theoretical model fit the empirical data. Good model fit is indicated by several parameters: the ratio of chi-square to degrees of freedom is less than 3 ($\chi^2/df < 3$) (Kline, 2016), comparative fit index is more than or equal to 0.90 (CFI $\geq$ 0.90) (Hu & Bentler, 1999), Tucker-Lewis Index is more than or equal to 0.90 (TLI $\geq$ 0.90) (Hu & Bentler, 1999), root mean square error of approximation is less than 0.06 to 0.08 with 90% confidence interval (RMSEA $<$ 0.06 to 0.08) (Schreiber et al., 2006), standardized root mean square residual is less than or equal to 0.08 (SRMR $\leq$ 0.08) (Hu & Bentler, 1999). To evaluate the standardized factor loading of each manifest variable, we used the criteria $\beta > |0.5|$, $p < 0.05$ (Gore et al. 2017; Khampirat, 2019).

Based on the results of the factor analysis that produced a fit model, the construct reliability of the high-stakes mathematics test was estimated. The composite reliability equation (Geldhof et al., 2014) was used to estimate construct reliability. Composite reliability indicates the internal consistency of indicators within a factor, and an acceptable level of reliability is 0.70 (Hair et al., 1999). In addition, the average variance extracted (AVE) was calculated for each factor to determine the proportion of variance explained by each factor in the proposed statistical model. Therefore, an acceptable AVE value is > 0.50 (Hair et al., 1999). All analyses in this study used the 'psych' package (Revelle, 2022) in RStudio (R Core Team, 2022).

FINDINGS AND DISCUSSION

Findings

In this section, descriptive statistics are presented to determine the characteristics of the data and confirm that the dataset is suitable for CFA. The conceptual model and fit statistics to confirm that the model fits the empirical data. Standardized factor loadings are also reported to examine each indicator's contribution to the latent variables. Finally, the results of the construct reliability analysis and the average variance extracted for each latent variable are presented.

Descriptive Statistics

The high-stakes mathematics test construct consists of four latent variables: number, algebra, geometry, and statistics. Each latent variable has a varying number of indicators. These indicators are referred to as manifest variables. Number consists of three manifest variables, algebra consists of six manifest variables, geometry consists of six manifest variables, and statistics consists of three manifest variables. A total of 18 manifest variables from the model

will be tested. The means, standard deviations, minimum values, and maximum values of the latent and manifest variables are presented in Table 1.

Table 1. Indicators of the High-Stakes Mathematics Test and Descriptive Statistics

Indicator	Abbreviation	<i>M</i>	<i>SD</i>	<i>Min.</i>	<i>Max.</i>
Numbers:	BIL	5.32	2.19	0	9
Number operations and their properties	BIL1	2.41	1.27	0	4
Ratios	BIL2	1.11	0.67	0	2
Sequences and series	BIL3	1.80	0.95	0	3
Algebra:	ALJ	6.16	2.96	0	11
Algebraic expressions	ALJ1	1.18	0.78	0	2
Linear inequalities in one variable	ALJ2	0.53	0.50	0	1
Sets and Venn diagrams	ALJ3	1.11	0.80	0	2
Relations and functions	ALJ4	0.94	0.78	0	2
Linear equations of a straight line	ALJ5	1.28	0.71	0	2
Systems of linear equations in two variables	ALJ6	1.11	0.76	0	2
Geometry:	GEO	9.65	3.18	2	16
Pythagorean theorem	GEO1	1.21	0.74	0	2
Quadrilaterals and their properties	GEO2	1.30	0.56	0	2
Triangles and their properties	GEO3	2.00	0.93	0	3
Relationships between lines and angles	GEO4	1.90	0.92	0	3
Elements of solid figures	GEO5	1.21	0.67	0	2
Surface area and volume of solid figures	GEO6	2.02	1.16	0	4
Statistics and Probability:	STAT	2.62	1.18	0	4
Measures of central tendency	STAT1	1.47	0.68	0	2
Data representation	STAT2	0.48	0.50	0	1
Sample points, sample space, and probability	STAT3	0.67	0.47	0	1

Construct Validity

To establish the construct validity of the high-stakes mathematics test, a second-order confirmatory factor analysis was conducted. The conceptual model is presented in Figure 1. To evaluate how well the conceptual model fits the empirical data, this study used several model fit indices, as shown in Table 2. Based on Table 2, all fit indices of the conceptual model (χ^2 , RMSEA, CFI, TLI, and SRMR) meet all goodness-of-fit model criteria. Thus, it can be concluded that the mathematical ability construct measured through four latent variables and 18 manifest variables fits the empirical data.

Table 2. Goodness of Fit Indices of Second-Order CFA of High-Stakes Mathematics Test

Fit indices	Good fit	Proposed model
χ^2	$p > 0.05$	$p = 0.134$
χ^2/df	$\chi^2/df < 3$	$(140.013)/(131) = 1.069$
RMSEA	$0 \leq \text{RMSEA} \leq 0.06$	0.015
CFI	$0.90 \leq \text{CFI} \leq 1.00$	0.985
TLI	$0.90 \leq \text{TLI} \leq 1.00$	0.983
SRMR	$0.00 \leq \text{SRMR} \leq 0.06$	0.041

As shown in Figure 1, most indicators on the first-order factor have moderate standardized loadings with their latent variables. On the latent variable of numbers (BIL), indicator BIL1 has the highest loading, and BIL2 has the lowest loading. In the algebra latent variable (ALJ), the ALJ1 indicator has the highest loading, followed by ALJ3, while the ALJ2 indicator has the lowest loading on the algebra latent variable. In the geometry latent variable (GEO), the GEO4 indicator has the highest standardized loading, followed by GEO6. Conversely, the GEO2 indicator has the lowest factor loading on the geometry latent variable. Furthermore, in the statistics latent variable (STAT), it was found that the STAT2 indicator had

the highest loading, while the lowest loading was contributed by the STAT1 indicator. Despite having varying values, all standardized loadings on the first-order factor were confirmed to be significant (see Table 3).

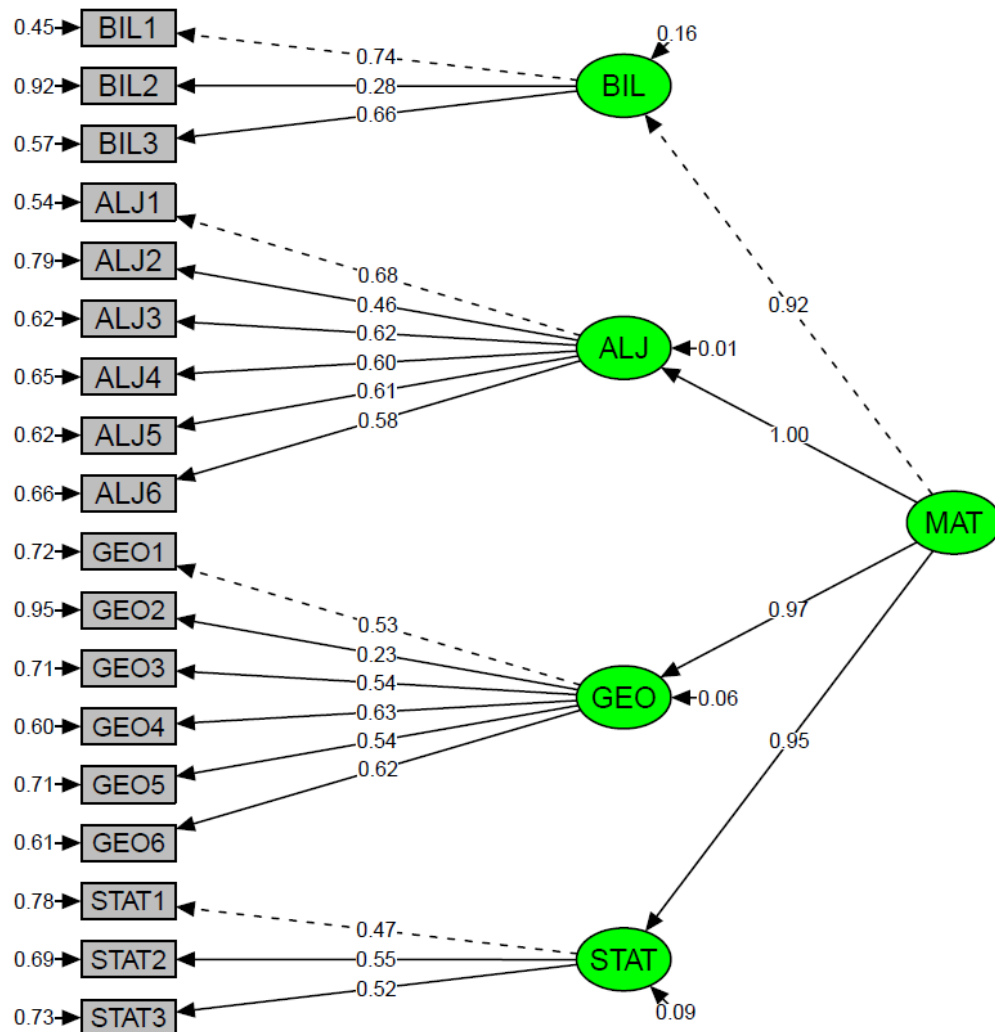


Figure 1. Second-Order Factor Model of High-Stakes Mathematics Test with Standardized Parameter Estimates

In Figure 1, all standardized factor loadings on the second-order factor were in the strong category. Algebra (ALJ) has the highest standardized loading factor of 1.00, and numbers (BIL) have the lowest loading factor of 0.92. All loading factors on the second-order factor are significant, as explained in the table. This indicates that the four latent variables in the model have a significant contribution to measuring mathematical ability.

The estimate column in Table 3 shows the unstandardized loading factors for each manifest variable and latent variable. In this study, only standardized factor loadings (Std. loading) are used to interpret the contribution of each variable. The Std. loading column in Table 3 shows that the loading factors of the manifest variables in the first-order factors range from 0.231 to 0.743. Meanwhile, in the second-order factors, the latent variables have loading factors ranging from 0.915 to 0.997. The $P(> |z|)$ column shows the significance of the loading factors of each variable in the first- and second-order factors. This column shows that all loading factors are significant. The Residual Column shows the error estimation values in the first- and second-order factors. The highest error is contributed by GEO2 (first-order), and the lowest is contributed by ALJ (second-order).

Table 3. Loading Factor, Residual, and R-Square of Second-Order Factor Model

Latent variable	Manifest Variable	Estimate	Std. Err	z-value	$P(> z)$	Std. loading	Residual	R square
Numbers (BIL)	BIL1	1.000				0.743	0.450	0.552
	BIL2	0.196	0.041	4.788	0.000	0.276	0.920	0.076
	BIL3	0.663	0.060	10.989	0.000	0.656	0.570	0.431
Algebra (ALJ)	ALJ1	1.000				0.675	0.540	0.456
	ALJ2	0.432	0.055	7.874	0.000	0.456	0.790	0.208
	ALJ3	0.934	0.087	10.72	0.000	0.617	0.620	0.380
	ALJ4	0.884	0.084	10.549	0.000	0.595	0.650	0.354
	ALJ5	0.822	0.074	11.141	0.000	0.615	0.620	0.378
	ALJ6	0.833	0.083	10.039	0.000	0.579	0.660	0.335
Geometry (GEO)	GEO1	1.000				0.529	0.720	0.280
	GEO2	0.332	0.094	3.521	0.000	0.231	0.950	0.054
	GEO3	1.284	0.155	8.307	0.000	0.536	0.710	0.287
	GEO4	1.482	0.166	8.931	0.000	0.629	0.600	0.396
	GEO5	0.925	0.116	7.96	0.000	0.535	0.710	0.286
	GEO6	1.858	0.217	8.577	0.000	0.624	0.610	0.389
Statistics (STAT)	STAT1	1.000				0.468	0.780	0.219
	STAT2	0.877	0.127	6.885	0.000	0.555	0.690	0.308
	STAT3	0.769	0.106	7.273	0.000	0.517	0.730	0.267
Mathematical Ability (MAT)	BIL	1.000				0.915	0.160	0.838
	ALJ	0.608	0.054	11.303	0.000	0.997	0.010	0.995
	GEO	0.436	0.050	8.748	0.000	0.969	0.060	0.939
	STAT	0.349	0.047	7.433	0.000	0.953	0.090	0.908

Note:

variances of residual BIL1, ALJ1, GEO1, STAT1, and BIL are fixed at zero so that they can be identified

In addition, the R-square column in Table 3 shows the total variance explained by each manifest variable. In the first-order factor, BIL1 explains the highest total variance (55.2%) compared to other manifest variables. Conversely, the total variance explained by GEO2 is the lowest (5.4%) compared to other manifest variables. The three indicators in the latent variable of numbers (BIL), namely BIL1, BIL2, and BIL3, can explain 7.6% to 55.2% of the total variance. In the latent variable of algebra (ALJ), the six indicators can explain 20.8% to 45.6% of the total variance. The six indicators in the latent variable of geometry (GEO) can only explain 5.4% to 39.6% of the total variance. Meanwhile, the three indicators in the latent variable of statistics (STAT) can explain 21.9% to 30.8% of the total variance. In the second-order factor, the latent variables BIL, ALJ, GEO, and STAT can explain 83.8% to 99.5% of the total variance. This indicates that the four latent variables, namely numbers, algebra, geometry, and statistics, are very satisfactory in explaining mathematical ability.

Construct Reliability

Table 4 presents the construct reliability (CR) and average variance explained (AVE) of the high-stakes mathematics test. The conceptual model in Figure 1 shows that there are five constructs of high-stakes mathematics: the main construct is mathematical ability (MAT), and the subconstructs are numbers (BIL), algebra (ALJ), geometry (GEO), and statistics (STAT). Of the four subconstructs, only ALJ has a CR > 0.70, while the reliability of the other three subconstructs is not very satisfactory (CR < 0.70). However, the main construct has very satisfactory reliability (CR = 0.98). The AVE of the four subconstructs is also not very satisfactory (AVE < 0.50), but the main construct has a very satisfactory AVE (AVE = 0.92). Therefore, the high-stakes mathematics test can explain 92% of the total variance.

Table 4. Construct Reliability and Average Variance Explained

Construct	CR	AVE
Mathematical Ability (MAT)	0.98	0.92
Numbers (BIL)	0.59	0.35
Algebra (ALJ)	0.76	0.35
Geometry (GEO)	0.69	0.28
Statistics (STAT)	0.52	0.26

Discussion

This study reveals that the construction of high-stakes mathematics tests for junior high school students in Indonesia is valid and reliable. There are four important factors for measuring mathematical ability, namely numbers, algebra, geometry and measurement, and statistics and probability. In addition, the 18 indicators contained in these factors make a significant contribution with varying loads. The proven construct validity indicates that the test measures the abilities or theoretical constructs to be measured (Allen & Yen, 1979; Nunnally, 1978; Retnawati, 2016) and that the test construction was developed based on an appropriate conceptual framework (Allen & Yen, 1979; Retnawati, 2016). The findings of this study indicate that four significant main components measure mathematical ability at the junior high school level, which is consistent with previous studies. (i.e., Bano et al., 2021; Kartianom & Mardapi, 2017; Rafi et al., 2023).

Furthermore, the findings indicate that the results of the second-order CFA analysis meet all indications of model fit, and all indicators contribute significantly to the factors. These results are consistent with the findings of previous studies (i.e., Apaydin, 2020; Khampirat, 2019; Suciati et al., 2022). First-order factor analysis shows that the factor loadings of all manifest variables are strong. This indicates that the constructs of numbers, algebra, geometry and measurement, and statistics and probability contribute strongly to mathematical ability. However, different results appear in the second-order factor results, where not all manifest variables are able to contribute strongly to the latent variables. The comparison and quadrilateral concepts have the lowest factor loadings compared to other indicators. This means that these two indicators contribute the least to the latent variables. Several studies report that comparison and quadrilateral material is difficult for junior high school students (Kartianom & Mardapi, 2017; Retnawati, Arlinwibowo, et al., 2017). Test participants experienced considerable difficulty in answering items on these materials, as reflected in the low mean scores for both indicators (see Table 1). Further studies are expected to investigate the impact of students' difficulty in answering items on an indicator on the magnitude of the loading factor in CFA.

In terms of reliability, the main construct (mathematical ability) has high reliability, but the reliability of the sub-constructs is not as high as that of the main construct. Within the framework of second-order CFA, our findings are consistent with those of previous studies by Khampirat (2019). However, these findings differ from those of Apaydin (2020). Based on these facts, it cannot be concluded that the reliability of the main constructs (first-order factors) in second-order CFA depends on the reliability of the sub-constructs (second-order factors). This issue provides an opportunity for other researchers to investigate the relationship between the reliability of main constructs and sub-constructs in the framework of second-order CFA.

In addition, the findings contribute to filling the gap in the literature on the validity and reliability of high-stakes mathematics tests. High-stakes mathematics tests, which have a significant impact on students, teachers, schools, and education stakeholders, cannot be separated from these two issues. This study seeks to reaffirm the importance of this issue and highlights the continuing need for further literature on the topic. In practical terms, the findings of this study can guide test developers in designing high-quality tests, especially in mathematics. This study emphasizes that establishing test validity cannot rely solely on expert judgments

regarding the feasibility and quality of test construction and item development. This procedure is often known as content validity. However, when tests have significant consequences for test takers, proving validity based on empirical data is crucial. Similarly, with regard to test reliability, test developers should not only use traditional approaches, but also try more advanced approaches in estimating the reliability of a test. In addition, our findings are also important for mathematics teachers. This study has shown which topics contribute significantly to students' mathematical abilities. This can be used by teachers to design effective mathematics learning.

Therefore, this study has several limitations. First, this study did not involve direct field data collection; therefore, the researchers were unable to control the situations and conditions under which the tests were administered. Second, although the sample used was taken randomly, the population only involved five provinces in Indonesia. In the context of Indonesia, which covers a vast area and consists of 38 provinces, the sample used is not representative enough to represent the diversity of student characteristics in Indonesia. Third, the construct tested in this study only focused on the mathematical content dimension, while the cognitive dimension was not tested. These limitations provide guidance for future studies. Future studies are expected to collect data directly so that researchers can ensure that the data collection process is conducted honestly and without manipulation. This is important to ensure the credibility of the data so that the results of the test validity are truly accurate. Although the literature does not recommend the use of very large sample sizes in CFA, it is very important to ensure that the sample is truly representative, especially for regions with highly variable population coverage. On this basis, we suggest that future studies should not focus on obtaining as many samples as possible, but rather on ensuring that all characteristics in the population are represented. Consistent with the procedures employed in this study, a sample size of 300 is considered sufficient for conducting CFA. In addition, future studies are expected to also consider cognitive dimensions or other attributes that may be inherent in the test. In other words, the analysis of the validity and reliability of a test's constructs in terms of cognitive and other dimensions is expected to strengthen the findings of this study and provide guidance for the development of standardized mathematics tests in the future.

CONCLUSION

This study concludes that the construct of the high-stakes mathematics test for junior high school level in Indonesia is valid. The mathematical ability measured through the test has four sub-constructs: numbers, algebra, geometry, and statistics. These four sub-constructs contribute significantly to explaining the mathematical ability of test participants. All indicators comprising the four sub-constructs of mathematical ability were also confirmed to make a significant contribution. This provides evidence that the test developed is valid for measuring the mathematical ability of test participants. In addition, the construct reliability and average variance explained by the test were also very satisfactory, proving that the test is reliable and can measure mathematical ability accurately. The findings of this study provide insight for mathematics test developers in determining key indicators in test development. These findings can also be used by educators to strengthen the mastery of mathematical content that has been proven to significantly contribute to students' mathematical abilities.

Authorship Contributions

AA: Conceptualization, Formal Analysis, and Writing-Original Draft; **MKM:** Conceptualization, Investigation, and Writing-Original Draft; **RR:** Supervision, Validation, and Writing-Review & Editing; **HR:** Supervision, Validation, Writing-Review & Editing, and Funding Acquisition; **EA:** Methodology, Validation, Software, and Writing-Review & Editing; **IR:** Methodology, Software, and Visualization; **MNR:** Investigation, Data Curation, and Visualization.

Funding Statement

This research does not receive funding.

Conflict of Interests

The authors declare that there are no potential issues of interest in this study.

Use of Generative-AI Tools Declaration

The authors declare that no generative artificial intelligence (AI) tools were used at any stage of this study or in the preparation of this manuscript, including idea development, data collection, data analysis, interpretation of results, and manuscript drafting.

Ethics Approval

This study did not require ethics approval because it involved a secondary analysis of anonymized student response data obtained with official permission from the authorized government institution responsible for national examination data in Indonesia. The dataset contained only participant identification codes, provincial codes, and dichotomously scored responses to test items, without students' names or school identifiers. The researchers had no access to personally identifiable information, and the data were analyzed solely for research purposes. Therefore, the study did not involve direct interaction with participants or the collection of identifiable personal information.

REFERENCES

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole Publishing Company.
- Apaydin, C. (2020). Second-order factor analysis of teachers' correct use of their voices scale in classroom management (TCUVS). *European Journal of Education Studies*, 6(11), 164–179. <https://doi.org/10.5281/zenodo.3627322>
- Bano, H., Khan, M. S., & Khan, S. I. (2021). Development of valid and reliable mathematics achievement test. *Ilkogretim Online - Elementary Education Online*, 20(2), 1642–1658. <https://ilkogretim-online.org/index.php/pub/article/view/2302>
- Demir, C. G., & Keleş, Ö. K. (2021). The impact of high-stakes testing on the teaching and learning processes of mathematics. *Journal of Pedagogical Research*, 5(2), 119–137. <https://doi.org/10.33902/jpr.2021269677>
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage.
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19(1), 72–91. <https://doi.org/10.1037/a0032138>
- Gore, J., Holmes, K., Smith, M., Fray, L., McElduff, P., Weaver, N., & Wallington, C. (2017). Unpacking the career aspirations of Australian school students: Towards an evidence base for university equity initiatives in schools. *Higher Education Research & Development*, 36(7), 1383–1400. <https://doi.org/10.1080/07294360.2017.1325847>
- Gregory, K., & Clarke, M. (2003). High-stakes assessment in England and Singapore. *Theory into Practice*, 42(1), 66–74. https://doi.org/10.1207/s15430421tip4201_9
- Hair, J. F., Anderson, R. E., Tahtam, R. L., & Black, W. C. (1999). *Multivariate data analysis* (5th ed.). Pearson Education.

- Heubert, J. P. (2000). High-stakes testing opportunities and risks for students of color, English-language learners, and students with disabilities. *National Center on Accessing the General Curriculum*.
- Heubert, J. P., & Hauser, R. M. (1999). High stakes: Testing for tracking, promotion and graduation. *National Academy of Sciences*.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Kartianom, K., & Mardapi, D. (2017). The utilization of junior high school mathematics national examination data: A conceptual error diagnosis. *REID (Research and Evaluation in Education)*, 3(2), 163–173. <https://doi.org/10.21831/reid.v3i2.18120>
- Kartowagiran, B., Munadi, S., Retnawati, H., & Apino, E. (2018). The equating of battery test packages of mathematics national examination 2013-2016. *SHS Web of Conferences*. <https://doi.org/10.1051/shsconf/20184200022>
- Khampirat, B. (2019). A second-order confirmatory factor analysis of the career aspirations scale of female Thai students. *International Journal of Learning, Teaching and Educational Research*, 18(11), 265–282. <https://doi.org/10.26803/ijlter.18.11.16>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
- Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy and Practice*, 19(1), 99–116. <https://doi.org/10.1080/0969594X.2011.604023>
- Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, 09(08), 2207–2230. <https://doi.org/10.4236/psych.2018.98126>
- Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6–14. <https://doi.org/10.1111/j.1745-3992.2004.tb00160.x>
- Marchant, G. J. (2004). What is at stake with high stakes testing? A discussion of issues and research. *Ohio Journal of Science*, 104(2), 2–7. <https://kb.osu.edu/items/84e45c5c-c662-5c06-bad4-84e803e91371>
- Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy*, 3(1), 82–100. <https://doi.org/10.2478/v10159-012-0004-x>
- Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, 36(8), 470–476. <https://doi.org/10.3102/0013189x07311608>
- Nichols, S. L., Glass, G. V., & Berliner, D. C. (2012). High-stakes testing and student achievement: Updated analyses with NAEP data. *Education Policy Analysis Archives*, 20(20), 1–35. <https://doi.org/10.14507/epaa.v20n20.2012>
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.
- R Core Team. (2022). R: A language and environment for statistical computing. <https://www.r-project.org/>
- Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., & Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-

- standardized school examination. *Pedagogical Research*, 8(1), em0145. <https://doi.org/10.29333/pr/12657>
- Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian (panduan peneliti, mahasiswa, dan psikometrian)*. Parama Publishing.
- Retnawati, H., Arliani, E., Apino, E., Djidu, H., Kartianom, N., Santoso, A., & Rasmuin, N. (2021). Mapping of difficult mathematics topics in vocational high schools based on national examination data. *Indonesian Journal of Educational Assessment*, 3(2), 107–127. <https://doi.org/10.26499/ijea.v3i2.83>
- Retnawati, H., Arlinwibowo, J., & Sulistyaningsih, E. (2017). The students' difficulties in completing geometry items of national examination. *International Journal on New Trends in Education and Their Implications*, 8(4), 28–41. <https://ijonte.elapublishing.net/makale/6544>
- Retnawati, H., Hadi, S., Munadi, S., Hadiana, D., Muhardis, M., Apino, E., Djidu, H., Rafi, I., Yusron, E., & Rosyada, M. N. (2019). When national examination no longer determining graduation, will students accomplish it seriously? *Indonesian Journal of Educational Assessment*, 2(2), 40–49. <https://doi.org/10.26499/ijea.v2i2.34>
- Retnawati, H., Hadi, S., Nugraha, A. C., Arlinwibowo, J., Sulistyaningsih, E., Djidu, H., Apino, E., & Iryanti, H. D. (2017). Implementing the computer-based national examination in Indonesian schools: The challenges and strategies. *Problems of Education in the 21st Century*, 75(6), 612–633. <https://doi.org/10.33225/pec/17.75.612>
- Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, 10(3), 257–276. <https://doi.org/10.12973/iji.2017.10317>
- Revelle, W. (2022). Package 'psych.' <https://cran.r-project.org/web/packages/psych/psych.pdf>
- Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, 99(6), 323–338. <https://doi.org/10.3200/JOER.99.6.323-338>
- Smith, M. L., & Fey, P. (2000). Validity and accountability in high-stakes testing. *Journal of Teacher Education*, 51(5), 334–344. <https://doi.org/10.1177/0022487100051005002>
- Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161–179. <https://doi.org/10.1080/00131880902891305>
- Stobart, G., & Eggen, T. (2012). High-stakes testing - value, fairness and consequences. *Assessment in Education: Principles, Policy and Practice*, 19(1), 1–6. <https://doi.org/10.1080/0969594X.2012.639191>
- Suciati, S., Munadi, S., & Sugiman, S. (2022). Estimation of test item parameters with polytomous item response using partial credit model (PCM). *Proceedings of the 2nd International Conference on Innovation in Education and Pedagogy (ICIEP 2020)*, 619, 232–237. <https://doi.org/10.2991/assehr.k.211219.042>
- Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics*. Pearson Education.