Southeast Asia Mathematics Education Journal
Volume 14, No 1 (2024)

Development of a Numeracy Assessment Instrument for Vocational High School (VHS) Students

[1]Jumini & [2]Kana Hidayati
[1]*State Vocational High School 1 Kalasan, Yogyakarta, Indonesia*
[2]*Department of Mathematics Education, Yogyakarta State University, Yogyakarta, Indonesia*

Abstract

Numeracy skills are essential components that must be mastered by Vocational High School (VHS) graduates to enter the workforce. Developing a quality instrument based on the Item Response Theory (IRT) approach is necessary to measure numeracy skills. Since the development process involved VHS students, the instrument is developed based on the characteristics of the users. This study aims to (1) develop a quality numeracy assessment instrument for VHS students and (2) describe the numeracy skill profile of VHS eleven graders in Sleman Regency based on the assessment result. The subjects of the wide-scale trial were 374 students from nine VHSs in the Sleman Regency of Indonesia, assessed in nine study areas. Data analysis used the Item Response Theory (IRT) of the Quest program with the outputs: (1) An instrument consisting of 35 questions involving multiple-choice, complex multiple-choice, short answer, and essay. All items are valid qualitatively based on expert judgment and quantitatively based on infit and outfit MNSQ value and unidimensionality. The test item reliability value is 0.96, including in the special category, and the person reliability estimate is 0.89, categorized as good. Based on Total Information Function (TIF) and Standard Error of Measurement (SEM), the instrument is reliable in estimating the numeracy skills of the students with the ability ranging from -3.2 to 3.3 logit or covering the ability categories of low, medium, and high. The difficulty level of all items is in the range of -2.0 to 2.0, falling into the good category. (2) Most students, covering 67% of VHS students in Sleman Regency, have numeracy skills at the Basic level.

Keywords: Assessment, Instrument, Numeracy, Vocational High School

Introduction

Education is one of the most vital aspects of national life, with the quality of a country's education reflecting the overall quality of other sectors and even determining the nation's future. As Tukiran (2020) states, the quality of a nation's education directly impacts the quality of its human resources, which are essential for national progress. In Indonesia, efforts to improve the quality of education and ensure equitable access focus on enhancing students' abilities and learning outcomes across all academic units under the Ministry of Education and Culture (Kemdikbud). Among the various competencies being promoted, numeracy is considered a key skill. According to the Indonesian Ministry of Education and Culture (Kemdikbud RI, 2020), numeracy involves the application of facts, concepts, procedures, and mathematical tools to solve problems in diverse contexts.

Several studies and research on numeracy highlight the critical importance of this skill. Doig et al. (2003: 13)) assert that numeracy skills are essential for meeting current and future demands. Similarly, Andreas Schleicher of the Organization for Economic Co-operation and

Development (OECD) emphasized that strong numeracy skills provide the best defence against poor health, unemployment, and low incomes (Kemdikbud RI, 2017b: 2). Furthermore, research by Jelatu, Mon, & San (2019) demonstrates a positive and significant relationship between numerical ability and the academic achievement of vocational high school (VHS) students. Consequently, improving numeracy skills is essential, and one effective strategy is through formal education. Supporting this, Green & Riddell (2012) which conclude that formal schooling is the dominant determinant of numeracy and problem-solving skills.

Regarding the numeracy abilities of Indonesian students, one essential assessment that measures these skills is the Program for International Student Assessment (PISA), organized by the Organization for Economic Co-operation and Development (OECD). Indonesia was one of 79 participating countries in 2018. This assessment evaluates the numeracy, reading, and science skills of students who have completed primary education or are approximately 15 years old. According to a report by the Indonesian Ministry of Education and Culture, the scores achieved by Indonesian students in PISA 2018 were lower than those of students in Peru and Brazil, two countries with similar characteristics to Indonesia. Moreover, Indonesian students' scores fell below the average for both ASEAN and OECD countries. The same report also revealed that around 71% of Indonesian students did not meet the minimum competency in numeracy, indicating that many struggle when faced with situations requiring mathematical skills to solve problems (Balitbang Kemdikbud, 2019: 50). This issue impacts students at the high school level, as foundational skills are critical for success in subsequent stages of education. These findings underscore the need for improvement in the numeracy skills of Indonesian students, which ongoing measurement and evaluation efforts must support.

Currently, numeracy assessments in Indonesia are conducted exclusively online and centrally by the Ministry of Education and Culture through the National Assessment (AN), which takes place once in the middle of the education level. However, the process could be more effective if educational institutions and educators also administered numeracy assessments independently, such as during mid-semester exams or other school-level evaluations. This assessment would allow schools and teachers to monitor students' progress in numeracy skills more frequently and efficiently. High-quality assessment tools that provide reliable and meaningful data must be developed to achieve this goal. These instruments would enable educational institutions to make informed decisions and take strategic steps to enhance the learning process and inform policy adjustments, ultimately improving students' numeracy skills.

The reality shows that quality numeracy assessment instruments have yet to be widely developed in Indonesia. However, a few examples can be found on the Indonesian Ministry of Education and Culture's website. Research on the development of such instruments still needs to be completed. For instance, there has been research on the development of numeracy instruments for junior high school students by Futri & Rosnawati (2022), the development of numeracy instruments for high school students by Pulungan (2014) using the Borg & Gall development model, and the creation of a mathematical literacy instrument for junior high school students by Suciati et al. (2020). Additionally, Suprawata (2022) developed a test

instrument to measure the numeracy skills of elementary school teachers, while Kurniawan et al. (2022) developed numeracy questions within the context of Javanese literature.

However, no specific numeracy assessment instrument has been developed for vocational high school (VHS) students. Given that VHS students are the intended users, involving them in the development process is essential to ensure the resulting instrument aligns with their characteristics. Furthermore, previous efforts to develop numeracy instruments have primarily used Classical Test Theory (CTT), rather than the more modern Item Response Theory (IRT). CTT has a notable weakness: the discriminant index and difficulty level of the items are dependent on the abilities of the students taking the test (Retnawati, 2015). Hambleton and Lord, as cited in Istiyono (2018), also pointed out several other drawbacks of CTT, including the dependence of test item statistics on the characteristics of the tested subjects, the high reliance of ability estimation on the test items, the uniform application of the standard error of score estimation to all test takers, and the lack of specific measurement error for each participant and item. Additionally, CTT only considers right or wrong answers without accounting for test-takers' answer patterns, and it is challenging to meet the assumption of parallel tests.

The weaknesses of Classical Test Theory (CTT) can be addressed by implementing Item Response Theory (IRT). IRT operates on two core principles. The first is the principle of relativity, which shifts the focus of measurement from the respondent or item to the respondent's performance on the item. This principle compares the respondent's ability and the item's difficulty level. If the respondent's ability is lower than the item's difficulty level, their answer will likely be incorrect. The second principle is probability. In IRT, a mathematical model predicts the likelihood that a respondent will answer an item correctly, depending on both their ability and the item's characteristics (Keeves & Alagumalai in Retnawati, 2014).

Based on the explanation above, it is clear that developing a high-quality numeracy assessment instrument for vocational high school (VHS) students is both essential and urgent. A reliable and accountable measurement tool is crucial to guide effective interventions that can enhance the numeracy skills of VHS students. The application of IRT sets this study apart, which allows for a more accurate analysis of item characteristics and student abilities. Unlike CTT, IRT ensures that the difficulty level of the items does not depend solely on the students' abilities. Instead, it provides a latent measure of student ability independent of item difficulty. This approach is expected to yield more reliable results, improving students' numeracy skills more effectively.

Method

This study is development research aimed at developing and validating educational products, especially test instruments designed to measure the numeracy skills of VHS students. The instrument set developed in this study includes an instrument grid, test items, answer keys, scoring guidelines, and item calibration results. The development model employed is a combination and modification of models proposed by Oriondo & Antonio (1998), Mardapi (2012), and Retnawati (2016). The stages of the development process are as follows: (1) Test planning, (2) Test writing, (3) Test validation through expert judgment, (4) Test trials, (5)

Analysis and revision, (6) Final test assembly, (7) Test implementation and (8) Interpretation of test results.

Instrument validity was assessed both qualitatively and qualitatively. Qualitative validity was established through expert judgment, while quantitative validity was measured using infit and outfit Mean Square (MNSQ) values and unidimensionality testing. Reliability of the instrument was estimated by calculating item reliability and person reliability. Additionally, the Total Information Function (TIF) and Standard Error of Measurement (SEM) were used to assess the consistency and accuracy of the instrument in measuring respondents' abilities or latent traits. The trial stage included a limited trial to evaluate the instrument's feasibility and a wide-scale trial to gather empirical evidence of its quality. The wide-scale trial involved 374 eleventh-grade students from nine VHSs in Sleman Regency of Indonesia, assessed in nine different areas of study. Data was collected through paper-based tests administered with direct and accountable supervision. The data was analyzed using the Quest program's Item Response Theory (IRT) approach, providing robust insight into the instruments' effectiveness.

Results and Discussion

Early Developing Stage

This stage begins with a study of conceptual and operational definitions to determine the content, context, and cognitive level involved in numeracy assessment. This study scrutinized various documents, the official website of the Indonesian Ministry of Education and Culture, and other relevant literature sources. This study is to construct a numeracy assessment instrument shown in the Table 1.

Table 1

The Construction of a Numeracy Assessment Instrument

Aspect	Component	Explanation
Content (scope or content of numeracy)	Number	Properties of sequence, representation, and number operations
	Algebra	Equations and inequalities, ratios and proportions, relations and functions (including number patterns)
	Geometry and Measurement	Introduction to plane shapes, use of surface area and volume concepts, understanding of measuring weight, length, time, discharge, volume, and area units with standard units
	Data and Uncertainty	Understanding, interpretation, and presentation of data and probability
Context (Background of the question)	Personal	Context related to students' activities in daily life
	Scientific	Contexts related to the application of mathematics to scientific activities and problems in science
	Socio-cultural	Contexts related to broader community or societal issues
Cognitive Level (the thinking process required to solve the problems presented)	Understanding (C2)	The thinking process requires understanding facts and procedures including the ability to interpret, give examples, group, summarize, draw conclusions, explain, and compare
	Application (C3)	The thinking process requires applying mathematical concepts in real and routine contexts
	Reasoning (C4, C5, C6)	The thinking process involves reasoning using mathematical concepts to solve various non-routine problems

Aspect	Component	Explanation
Question Form	Multiple-choice	Only one correct answer
	Complex Multiple-choice	Several correct answer
	Short Answer	Only the final answer
	Essay	Requires structured answer

In this stage, the instrument's qualitative validity, as judged by expert judgment, proved that all items were valid in content. However, several revisions were suggested. The validation result shows that the developed test is relevant to the indicators, contents, contexts, and users' cognitive level, thus appropriate to measure what should be measured, namely the numeracy abilities of VHS students.

Trial Stage

After the validation stage, the instrument was tested on a limited sample. This trial aimed to evaluate the instrument's feasibility and collect corrections and input from the point of view of the instrument users, a limited number of students and teachers. The evaluation in this limited trial is a qualitative evaluation dealing with the effectiveness and clarity of language, the function of symbols/images/graphics, and the suitability of the problem/context used. This limited trial produced several essential points as the basis for improvement, including reducing the instrument items from 47 to 40 based on considerations of the available time allocation, the length of text in the questions, and the difficulty level of the questions categorized as medium to high based on the students' and teachers' points of view. Item reduction is carried out by considering the representation of indicators, content, and form of questions of the test that will be tested further. In addition to reducing the number of items, revisions were made regarding sentence structure, pictures, graphics, symbols, and contexts, and items were replaced as needed.

The next stage is a wide-scale trial with respondents involving 374 students from nine VHSs in Sleman Regency, Indonesia, from various fields of study. The responses collected from this trial were then analysed using the Rasch model by the Quest program to prove the empiric quality of the instrument. This analysis includes proving validity using infit and outfit MNSQ and unidimensionality, estimating item and person reliability, estimating TIF and SEM, and estimating difficulty level and student ability.

Data Analysis

Proving the Assumptions of IRT Approach

The analysis using the IRT approach was started by proving the assumptions of IRT, namely unidimensionality, local independence, and parameter invariance. The unidimensionality was proven by figuring out the Eigenvalue from the Total Variance Explained table, which resulted in 11 factors with Eigenvalues higher than one that could be extracted. The first factor was the most dominant because it explained 23.131% of the Variance with an Eigenvalue of 9.021. The Eigenvalue of the first factor was almost four times that of the second factor. According to Naga (1992: 297), the unidimensional condition is fulfilled if the Eigenvalue of the first factor

is multiple times greater than those of the second factor and the following factors. It indicates that the instrument measures only one dimension, the numeracy dimension. In addition to fulfilling the IRT assumptions, unidimensionality is also used to prove instrument validity. Instrument unidimensionality is important to signify whether the instrument developed can measure what should be measured. (Sumintono & Widhiarso, 2014).

The second assumption is local independence. There are two types of local independence: local independence of the items and local independence of student responses. Local independence of the test items indicates that a student's ability to answer an item does not affect his ability to answer another item. It implies that the probabilities of a student correctly answering an item and other items do not affect each other (Purnama & Alfarisa, 2020). The unidimensionality can prove this local independence (Mars, 2010 in Retnawati, 2014: 8). Meanwhile, the covariance matrix can prove local independence in student responses. First, students were divided into ten groups based on their abilities. As a result of the analysis, they used the Quest Program with a logit scale. The covariance matrix obtained with the help of Excel is shown in Table 2.

Table 2

The Covariance Matrix

	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
K1	0.38	0.10	0.03	0.02	0.04	0.03	0.02	0.03	0.04	0.08
K2		0.04	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.03
K3			0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
K4				0.00	0.00	0.00	0.00	0.00	0.00	0.01
K5					0.01	0.00	0.00	0.01	0.01	0.01
K6						0.00	0.00	0.00	0.00	0.01
K7							0.00	0.00	0.00	0.01
K8								0.01	0.00	0.01
K9									0.01	0.01
K10										0.03

Based on Table 2**,** the covariance among groups shown by the elements outside the main diagonal is close to zero. According to Hambleton and Swaminathan (1985), a value close to zero demonstrates that the assumption of local independence is proven. It indicates that the correct response of a student over an item does not affect that of other students. This argument is supported by Ojerinde (2013) that argues that local independence does not mean that the items are not correlated, but its performance is independent and depends on students' abilities.

The third assumption is parameter invariance assumption: ability parameter invariance and items parameter invariance. Item parameter invariance constitutes that the level of a student's ability to respond an item does not change the item parameter. Meanwhile, ability parameter invariance refers to a student's latent trait or ability that the items' difficulty level will not influence (Retnawati, 2014: 3). The proof of item parameter invariance was intended to estimate the difficulty level since the Rasch model was employed. The measure of the difficulty level of all items used the response data from two groups of students and produced a pair of difficulty levels for each item shown in Figure 3.

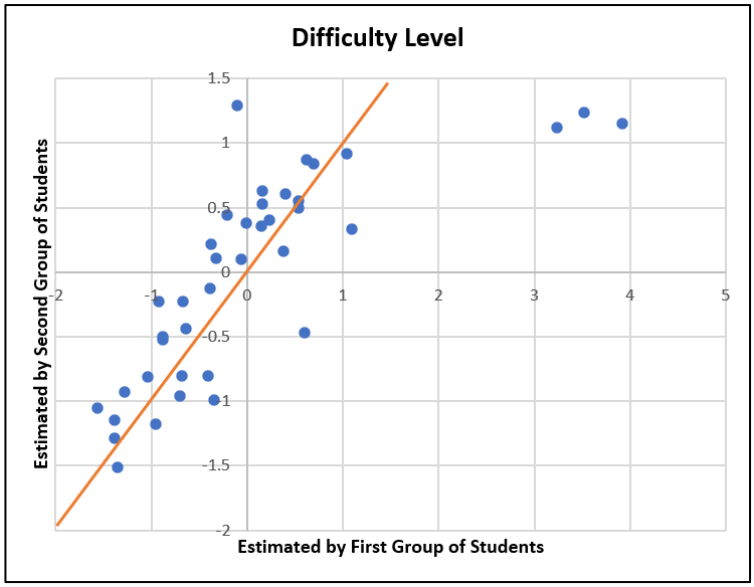

Figure 3. The Scatter Plot of Difficulty Level Invariance

Figure 3 shows that the plot distribution is close to the $y = x$ line which means that the two groups of students with different abilities produced almost similar estimates of difficulty level. Thus, the difference in the students' abilities does not change the item's difficulty level and verified the difficulty level invariance of the developed instrument.

The ability invariance was assessed by estimating the ability of all students using two groups of items to produce a pair of abilities for each student. The ability estimation was carried out using the R program for each data. Two pairs of student's abilities produced are demonstrated in the scatter plot in Figure 4.

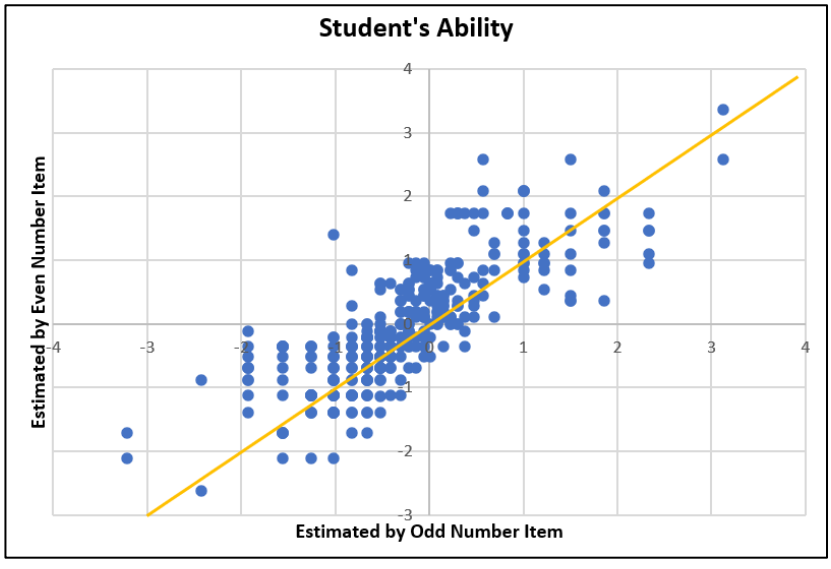

Figure 4. The Scatter Plot of Ability Invariance

Figure 4 shows that the plots of a student's abilities estimated by two groups of items are identical because the plots approach the $y = x$ line which means the student's abilities remain

the same although they do the items with different difficulty levels. Thus, the ability invariance is proven.

Fit Test Model

After testing the IRT assumptions, the parameter estimates for all items based on the response data of all students can be interpreted. An analysis of the item characteristics was carried out with the help of the Quest Program. This program can analyse dichotomous, polytomous, or a combination of dichotomous and polytomous data. The dichotomous data were analysed using the Rasch IRT approach. In contrast, the polytomous data were analysed using the Partial Credit Model (PCM), in which the PCM itself is an extension of the Rasch model for response data with more than two categories. The Quest outputs are the statistical fit or model fit, difficulty level, and students' abilities. The data analysed at this stage were the responses of 40 instrument items proven valid in content and constructs, consisting of 21 multiple-choice items, six complex multiple-choice items, six short answer items and six essay items.

The model fit test in this study is based on the Infit MNSQ value, one of the Quest Program outputs. The model fit test is not based on the chi-square value because according to Hooper, Couglan, and Mullen (2008: 57), the chi-square is sensitive to sample size, so it is likely to be rejected if the sample size is large. The model fit test or item fit explains the extent to which the sample pattern of response to an item is consistent with the responses of others to other items (Wright & Stone. 1999: 66-82). Item fit follows the rule that the Item Characteristic Curve (ICC) will be flat if the INFIT MNSQ magnitude is higher than 1.30 or less than 0.77 (Keeves & Alagumalai, 1999: 36). This study corroborates the argument of Aiken and Khoo (1996: 30 & 90) that states that an item fits the model if the MNSQ Infit value ranges from 0.77 to 1.30. Quest output related to item fit based on the Infit MNSQ value in this research proves that from a total of 40 items analysed with the Rasch/PCM model, 35 items fit the model.

Instrument Validity

In addition to the model fit test. the infit MNSQ value along the outfit MNSQ was also used to prove the instrument validity because model fit correlated with item validity. According to Reise (1990), the item's validity must be questioned if it does not fit with the model because it is likely that the item measures the ability that is not intended to be measured. The output of the Quest program that can be used as a criterion for item validity is the infit MNSQ value, which ranges from 0.77 to 1.3 (Setyawarno, 2017), and the outfit MNSQ value in the range of 0.5 to 1.5 (Sumintono and Widhiarso, 2015). Thirty-five questions were proven to fit the Rasch model because the infit MNSQ and outfit MNSQ values are within the criteria above. Thus, based on the expert assessment stage, infit and outfit MNSQ values, and unidimensionality, the 35 questions developed are valid.

Instrument Reliability

The subsequent analysis was conducted on instrument reliability based on the estimates of item reliability and person reliability as the outputs of Quest. The item and person reliability estimation values reached 0.96 and 0.89, respectively. According to Perdana (2018), the reliability of the Rasch model falls into the Special category if exceeding 0.94 and the Good category if ranging from 0.81 to 0.90. Based on these criteria, the item reliability of this instrument belongs to the Special category, meaning that almost all items fit the Rasch model. Furthermore, the person reliability falls into the Good category, which means the consistency of respondents in answering questions is also good, or respondents need to be more careful in answering. Thus, the instrument can provide reliable results.

The Total Information Function (TIF) and the Standard Error of Measurement (SEM) were also used to complement the instrument's strength. The information function can reveal the strength of an instrument at both the item level and the whole level. The strength in this context is the strength to measure the latent ability of the students (Myszkowski, 2019). The item information function establishes the strength of the information function of a set of instruments, referred to as the Total Information Function (TIF) which is the sum of the information functions of all its constituent items. The performance of the TIF and SEM numerical instruments in this study is presented Figure 5.

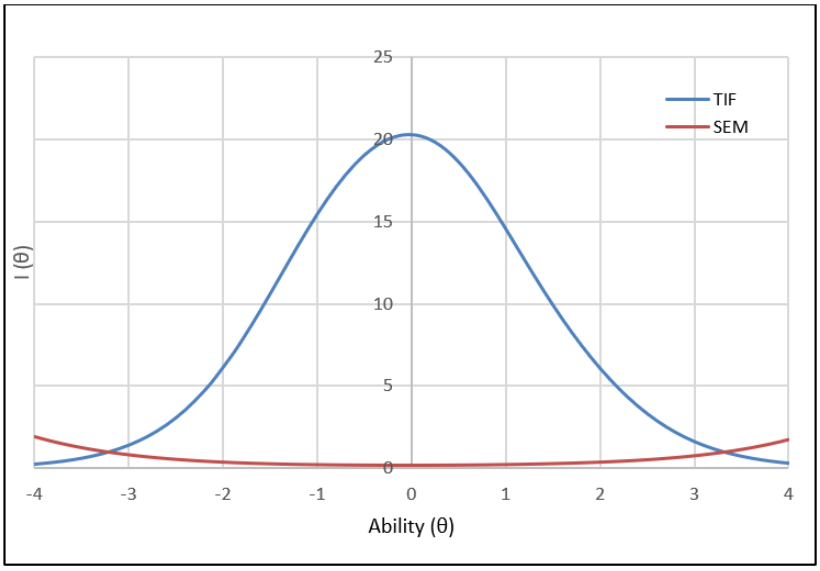

Figure 5. Curves of TIF and SEM

Figure 5 shows that the maximum test information function score represented by the blue curve is 20.269 at an ability (θ) of 0.0 and SEM of 0.222. This indicates that the test works optimally when applied to students with an ability of 0.0 logit scale or in the medium category. Hambleton (Wiberg, 2004) states that if the instrument has a TIF score higher than 10, it is reliable. The curves of TIF and SEM intersect at the values of θ are -3.2 and 3.3, hence, the instrument is reliable for measuring students' numeracy skills at low, medium, and high abilities.

Difficulty Level

The item parameter estimated in this analysis is the item difficulty level (threshold) because the model used is Rasch. An item is in the Good category if the difficulty level is between -2.0 and 2.0 on the logit scale (Hambleton & Swaminathan, 1985a: 36). The difficulty level range is made more specific by categorizing it into five difficulty levels. Table 3 summarizes the estimated difficulty level parameters of the 40 items and their categories.

Table 3
Results of Difficulty Level Estimation

Category	Difficulty Level (b)	Amount	Percentage (%)
Very Difficult	$b > 2$	0	0.00
Difficult	$1 < b \leq 2$	5	12.82
Moderate	$-1 < b \leq 1$	30	76.92
Easy	$-2 \leq b \leq -1$	5	12.82
Very Easy	$b < -2$	0	0.00
TOTAL		40	100.00

Referring to Table 3, there were no items outside the range of -2.0 to 2.0, so 35 valid items are also good in terms of difficulty level. In addition, the distribution or proportion of the difficulty level is balanced and meets the normal curve equilibrium where the items belonging to the medium category cover the most significant proportion. Moreover, the items in the Easy and Difficult categories are almost equal. This proportion is subject to the primary reference for preparing test items as Sudjana (2012: 135) states that the curve balance among the item difficulty levels is one of the essential references for the good development of a test instrument.

Student Numeracy Abilities

The estimated ability of 374 students obtained from the Quest output is presented on a logit scale of -4.0 to 4.0. The categorization of students' numeracy abilities was set based on the normal distribution and presented in the form of an appropriate diagram. The reason for using the normal distribution is that it is closely related to the purpose and follow-up of the numeracy assessment. The numeracy assessment aims to reveal the numeracy ability of each student which can further be used to improve learning strategies at the school level, instead of doing individual remedial, so that it is not the minimum cut of the score needed, rather, it is grouping the students' numeracy abilities.

To facilitate the presentation and interpretation, the ability level in this logit scale of -4.0 to 4.0 is converted into a standard scale of 0 to 100, as the common values known in education, especially in Indonesian schools, calculated using the equation follows (A is ability in standard scale and θ is ability in logit scale).

$$A = \frac{100}{8}(\theta + 4)$$

The converted numeracy ability levels were then grouped into four categories as suggested by the Centre for Educational Assessment of the Ministry of Education and Culture

(Pusmendik Kemdikbud), including Advanced, Proficient, Basic, and Special Intervention Required (Kemdikbud RI, 2020). The categories of the students' numeracy abilities are presented in Table 4.

Table 4

Student Numeracy Ability

Numeracy Ability (X)	Category	Number of students	Percentage (%)
$X \geq \bar{x} + 1.5\ SD$	Advance	11	2.94
$\bar{x} \leq X < \bar{x} + 1.5\ SD$	Proficient	65	17.38
$\bar{x} - 1.5\ SD \leq X < \bar{x}$	Basic	252	67.38
$X < \bar{x} - 1.5\ SD$	Special Intervention Required	46	12.30
Total		374	100

$\bar{x}$ = Ideal mean
SD = Ideal standard deviation

Table 4 **s**hows that the numeracy ability of VHS students in Sleman Regency that was successfully captured through this study is mostly still at the Basic level (67.38%). According to the Pusmendik Kemdikbud RI, students at this level have basic mathematical skills involving the computation of direct equations, basic concepts related to geometry and statistics, and solving simple, and routine mathematical problems. In addition, at this level, students still need help applying basic concepts to relevant situations (Kemdikbud RI, 2020). The follow-up needed by students at this level is the provision of examples of applying essential concepts and joint discussion to interpret and draw conclusions on problem-solving (Kemdikbud RI, 2020).

Conclusion

The conclusions regarding this research are as follows:

1. This research produces a set of numeracy assessment instruments for VHS students that are theoretically and empirically proven to be good quality. These instruments assess students' ability in numbers, algebra, geometry, measurement, data, and uncertainty in scientific, personal, and sociocultural contexts.
2. The cognitive levels measured include understanding, application, and reasoning. The instrument consists of 35 items: multiple-choice, complex multiple-choice, short answer, and essay. All the items are proven valid qualitatively and quantitatively, reliable and have good difficulty levels.
3. The measurement of the numeracy skills of VHS students in Sleman Regency, Indonesia, shows that most of the students, 67.38%, are still at the basic level.

Recommendation

Based on this research's conclusions, the recommendations are as follows.

1. Regarding Product Utilization

   Because this instrument has been proven to be of good quality, it can be directly used for numeracy assessment by teachers, schools, and other interested groups with the necessary adjustments.

2. Regarding Instrument Repair
   This instrument needs to be retested with stricter supervision to ensure its better quality. In addition to supervision, the number of test subjects needs to be increased to ensure the adequacy of the sample for each item and increase the number of parameters that can be estimated.

3. Regarding Further Development

   Developing an instrument equivalent to this instrument is needed to support research in numeracy, experimental research, developmental research, correlational research, and other types of research that will contribute to education, especially in increasing student numeracy.

Acknowledgment

Authors gratefully acknowledge funding from Indonesia Endowment Fund for Education or Lembaga Pengelola Dana Pendidikan (LPDP), Indonesian Ministry of Finance.

Reference

Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system version 2.1*. The Australian Council for Educational Research.

Balitbang Kemdikbud. (2019). *Laporan Nasional PISA 2018 Indonesia*. Kemdikbud RI.

Doig, B., McCrae, B., & Rowe, K. (2003). *A Good Start to Numeracy*. COmmonwealth of Australia.

Futri, V. I., & Rosnawati, R. (2022). *Pengembangan Instrumen Penilaian Kemampuan Numerasi Peserta Didik SMP pada Pembelajaran Matematika* [Universitas Negeri Yogyakarta]. https://eprints.uny.ac.id/73262/

Green, D. A., & Riddell, W. C. (2012). Understanding Educational Impacts: The Role of Literacy and Numeracy Skills. *11th IZA/SOLE Transatlantic Meeting of Labor Economists*.

Hambleton, R. K., & Swaminathan, H. (1985a). *Item Response Theory Principles and Aplications*. Kluwer Nijhoff Publishing.

Hambleton, R. K., & Swaminathan, H. (1985b). *Items Response Theory: Principles and Application*. Kluwer-Nijjhoff Publish.

Hooper, D., Couglan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, *6*(1), 53–60. https://www.researchgate.net/publication/254742561_Structural_Equation_Modeling_Guidelines_for_Determining_Model_Fit

Istiyono, E. (2018). *Pengembangan Instrumen Penilaian dan Analisis Hasil Belajar Fisika Dengan Teori Tes Klasik dan Modern* (First). UNY Press.

Jelatu, S., Mon, E. M., & San, S. (2019). Relasi Antara Kemampuan Numerik Dengan Prestasi Belajar Matematika. *Lectura: Jurnal Pendidikan*, *10*(1). https://doi.org/10.31849/lectura.v10i1.2390

Keeves, J. P., & Alagumalai. (1999). *New Approach to Measurement*. Pergamon, An Imprint of Elsevier Science.

Kemdikbud RI. (2017). *Materi Pendukung Literasi Numerasi*. Kemdikbud RI.

Kemdikbud RI. (2020). *AKM dan Implikasinya pada Pembelajaran*. Pusat Asesmen dan Pembelajaran Badan Penelitian dan Pengembangan dan Perbukuan Kementerian Pendidikan dan Kebudayaan Republik Indonesia.

Kurniawan, A. P., Budiarto, M. T., & Ekawati, R. (2022). Pengembangan Soal Numerasi Berbasis Konteks Nilai Budaya Primbon Jawa. *JRPM: Jurnal Review Pembelajaran Matematika*, *7*(1), 20–34. https://doi.org/10.15642/jrpm.2022.7.1.20-34

Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Nuha Medika.

Myszkowski, N. (2019). Development of the R library "jrt" : Automated Item-Response Theory procedures for judgment data and their application with the Consensual Assessment Technique. *Psychology of Aesthetics Creativity and The Arts*. https://doi.org/10.1037/aca0000287

Naga, D. S. (1992). *Pengantar Teori Skor Pada Pengukuran Pendidikan*. Penerbit Gunadharma.

Ojerinde, D. (2013). *Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of The Comparability of Item Analysis Result*. Semantic Scholar. https://www.semanticscholar.org/paper/Classical-Test-Theory-(-CTT-)-VS-Item-Response-(-)-Ojerinde/aa27637d699b7e26ee1f7578d0fe2b8173ad769f

Oriondo, L. L., & Antonio, D. E. M. (1998). *Evaluation Educational Outcomes*. Rex Printing Compagny.

Perdana, S. A. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48.

Pulungan, D. A. (2014). Pengembangan Instrumen Tes Literasi Matematika Model PISA. *Journal of Educational Research and Evaluation*, *3*(2).

Purnama, D. N., & Alfarisa, F. (2020). Karakteristik Butir Soal Try Out Teori Kejuruan Akuntansi SMK Berdasarkan Teori Tes Klasik dan Teori Respons Butir. *Jurnal Pendidikan Akuntansi Indonesia*, *18*(1), 36–46. https://doi.org/10.21831/jpai.v18i1.31457

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137. https://doi.org/10.1177/2F014662169001400202

Retnawati, H. (2014). *Teori Respons Butir dan Penerapannya*. Parama Publishing.

Retnawati, H. (2015). Perbandingan Estimasi Kemampuan Laten antara Metode Maksimum Likelihood dan Metode Bayes. *Jurnal Penelitian Dan Evaluasi Pendidikan, 19*(2). https://doi.org/10.21831/pep.v19i2.5575

Retnawati, H. (2016). *Validitas Reliabilitas & Karakteristik Butir*. Parama Publishing.

Setyawarno, D. (2017). *Upaya Peningkatan Kualitas Butir Soal dengan Analisis Aplikasi quest*. Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Yogyakarta.

Suciati, Munadi, S., Sugiman, Ratna, W. D., & Febriyanti. (2020). Design and Validation of Mathematical Literacy Instruments for Assessment for Learning in Indonesia. *European Journal of Educational Research*, *9*(2), 865–875. https://doi.org/10.12973/eu-jer.9.2.865

Sudjana, N. (2012). *Penilaian hasil Proses Belajar Mengajar*. Remaja Rosdakarya.

Sumintono, B., & Widhiarso, W. (Eds.). (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial (Edisi Revisi)*. Trim Komunikata Publishing House.

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

Suprawata, I. G. (2022). *Pengembangan Tes Numerasi Guru (Tugu) untuk Mengukur Kemampuan Numerasi Guru SD dengan Konteks Lingkungan Sekolah* [Universitas Pendidikan Ganesha]. https://repo.undiksha.ac.id/10663/

Tukiran, M. (2020). *Filsafat Manajemen Pendidikan*. Kanisius.

Wiberg, M. (2004). *Classical Test Theory vs. Item Response Theory*. Umea Universitet.

Wright, B., & Stone, M. (1999). *Measurement Essentials* (2nd ed.). Wide Range.