Issue VI, 22 November 2023
e-ISSN 2707-9481 Institute of Metallurgy and Ore Beneficiation, Satbayev University, Almaty, Kazakhstan
ISBN 978-601-323-356-7 https://doi.org/10.31643/2023.22

Alan Rifqi Kamal
Department of Educational Research
and Evaluation, Graduate School,
Yogyakarta State University (UNY), Indonesia
E-mail: alanrifqi.2021@student.uny.ac.id

Edi Istiyono
Department of Educational Research
and Evaluation, Graduate School,
Yogyakarta State University (UNY), Indonesia
E-mail: edi_istiyono_uny@yahoo.co.id

Analysis of numeracy ability test item characteristics grade VIII students with mixed model item response theory (IRT) approach

Abstract: Basic knowledge of mathematics is essential for solving problems contextually. Mathematics has a function for the development of the ability to calculate, measure, find, and use mathematical formulas that can provide students with an understanding of concepts related to life phenomena. One ability that is synonymous with understanding problems contextually is numeracy ability. Numeration has a main focus, namely the ability of students to formulate, apply, and be able to interpret mathematics in various contexts that include mathematical reasoning and using mathematical concepts, methods, facts, and auxiliary media, explaining, and predicting phenomena in everyday life. This study aims to determine the construct of numeracy ability test instruments for class VIII public junior high school students in Pekalongan Regency, determine the quality of numeracy ability test instruments for class VIII public junior high school students in Pekalongan Regency, and determine the numeracy ability profile of class VIII public junior high school students in Pekalongan Regency. This research method approaches quantitatively by developing instruments using CFA and IRT mixed models. This research was conducted at the junior high school level within the scope of the education office of Pekalongan Regency, Central Java Province, by taking 6 schools as samples. Content validity using Aiken V and Cronbach Alpha reliability as well as item characteristics with mixed IRT and descriptive analysis. The results of this study, namely (1) Construction of numeracy ability instruments for grade VIII State Junior High School students, which are related to the content of algebra, numbers, geometry, and measurement, as well as data and uncertainty. In addition, using personal, socio-cultural, and scientific contexts, using cognitive levels of understanding, application, and reasoning, (2) The quality of numeracy ability instruments is declared valid and reliable, and in construct validity all items are fit as seen from the *Loading Factor Standardized Solution* value of more than 0.3 and *p-value* < 0.05 and the reliability of the high category and the estimated characteristics of the items show that the question items are included in the category both in terms of difficulty, and (3) The numeracy ability of junior high school students in Pekalongan Regency shows that there are 36 students out of 599 students classified as proficient with a percentage of 6%, 139 students out of 599 students classified as proficient with a percentage of 23%, 390 students out of 599 students classified as basic with a percentage of 65%, and 34 students out of 599 students classified as needing special intervention with a percentage of 6%.

Keywords: Item Characteristics, Numeracy Ability, Mixed Model Item Response Theory (IRT).

Cite this article as: Alan Rifqi Kamal, Edi Istiyono (2023). Analysis of numeracy ability test item characteristics grade VIII students with mixed model item response theory (IRT) approach. *Challenges of Science*. Issue VI, 2023, pp. 184-195. https://doi.org/10.31643/2023.22

Introduction

Learning mathematics is not just about understanding concepts and theorems in mathematics, but students must be able to use logical reasoning to be able to solve problems or predict phenomena. Teacher involvement has an important role in achieving the success of the learning process. Teachers have a role in helping students to understand mathematical concepts. In this process students can use mathematics in solving contextual problems in accordance with mathematical concepts (OECD, 2019). Mathematics has a close relationship with everyday life problems, especially in terms of the ability to calculate, measure, and find patterns.

This is in line with the objectives of mathematics learning as outlined in the Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 21 of 2016 concerning Content Standards for Primary and Secondary Education, namely: 1) understanding mathematical concepts, describing the relationship between mathematical concepts and applying concepts flexibly, accurately, efficiently, and precisely in solving problems, 2) utilizing reasoning patterns of nature from mathematics, develop or manipulate mathematics in making general conclusions from an event, formulating evidence, or describing the results of mathematical thoughts and statements, 3) solving problems including determining known elements, assembling mathematical solving models, processing mathematical models, and providing appropriate solutions, and 4) communicating the results of thoughts or ideas with diagrams, tables, symbols, or other supporting media in order to clarify problems or circumstances. Basic skills for students need to exist so that the goals of national education are achieved. The basic ability to learn mathematics is closely related to numeracy ability.

The Ministry of Education and Culture (Kemendikbud) is very concerned about the urgency of numeracy ability. This can be seen from Indonesia's participation in *the Program for International Student Assessment* (PISA) driven by the *Organization for Economic Cooperation and Development* (OECD) which measures mathematical literacy skills which shows the results of PISA rankings obtained by Indonesia from 2003 to 2018; Indonesia ranked 38 out of 40 OECD member countries in 2003, Indonesia ranked 50 out of 57 OECD member countries in 2006, Indonesia ranked 61 out of 65 OECD member countries in 2009, Indonesia ranked 64 out of 65 OECD member countries in 2012, Indonesia ranked 69 out of 76 OECD member countries in 2015, Indonesia ranked 72 out of 78 OECD member countries in 2018 (OECD, 2019). These results show that there is still low mathematical literacy ability in Indonesia.

Basic knowledge of mathematics is essential for solving problems contextually. Mathematics has a function for the development of the ability to calculate, measure, find, and use mathematical formulas that can provide students with an understanding of concepts related to life phenomena (Megawati &; Sutarto, 2021). Learning mathematics is not just about understanding concepts, but can apply concepts that have been understood to solve problems. One ability that is synonymous with understanding problems contextually is numeracy ability.

Numeration means the ability of students to use their reasoning. Numeration has a main focus, namely the ability of students to formulate, apply, and be able to interpret mathematics into various contexts that include mathematical reasoning and use mathematical concepts, methods, facts, auxiliary media, explain, and predict phenomena in everyday life (Puspaningtyas &; Ulfa, 2020). It is important for students to understand numeracy which can later help students understand the role or benefits of mathematics in everyday life.

Numeracy ability is the ability or ability of students in terms of utilizing various kinds of numbers, diagrams, tables, symbols, or other supporting media that have a connection with mathematics to solve contextual problems and solve the information presented then be able to interpret the results of the analysis to predict and make decisions. More concisely, revealed (Ministry of Education and Culture, 2020), numeracy ability is the ability to think using concepts, procedures, facts, and mathematical media to solve everyday problems in various types of contexts. Numeration has a meaning as the ability of students to understand and use their mathematical knowledge in explaining phenomena, solving problems, or determining decisions in everyday life. This is in accordance with research conducted by (Braak &; Størksen, 2021) which explains that mathematical numeracy skills based on experience, discovery, experiments, or observations that have been made are still developing cumulatively. In everyday life we very often encounter phenomena related to numeracy skills such as when shopping, calculating height or weight, determining drug doses, regulating diet and nutrition, and many more related to student numeracy.

The Indonesian government considers the importance of numeracy skills for students to train students' reasoning in daily activities. This is done by replacing the UN by the Ministry of Education and Culture per 2021 to the Minimum Competency Assessment (AKM) in order to prepare students who have skills in the 21st century will be carried out a fundamental competency assessment to measure the ability to reason using mathematics or numeracy (Ministry of Education and Culture, 2020). Numeracy skills need to be possessed by students as a fundamental ability to be able to apply mathematical concepts in everyday life.

In order to find out the results of numeracy that is able to explain students' abilities according to their conditions, an accurate measuring instrument is needed. Assessment in mathematics subjects is used to measure students' numeracy abilities related to students' basic knowledge, namely the ability to apply, and process the understanding of mathematical concepts into the phenomena obtained which are divided into 4

main points, namely numbers, algebra, measurement and geometry, as well as data and uncertainty (Ministry of Education and Culture, 2020). In order to find out students' numeracy skills, measuring instruments in the form of test instruments are needed as a way of collecting data. Instruments used to collect information through student answers as evidence of learning outcomes which then the results are used to determine student characteristics (Istiyono, 2020; Retnawati, 2016). In this case it is necessary to use a measurement model called *Item Response Theory* (IRT).

Item Response Theory (IRT) is widely used in test analysis in educational, psychological, and using probabilistic models (Gunawan *et al*, 2020; Mardapi, 2017). The mathematical model means that students have the opportunity to answer question items correctly depending on student abilities and item characteristics (Retnawati, 2014). This means that students who have high abilities will have a greater chance of answering questions correctly than students who have low abilities. In addition, there are three assumptions that must be met in *Item Response Theory (*IRT): (1) unidimensional, meaning that each question item measures only one ability, (2) local independence, meaning that there is no correlation between test taker responses to different questions, and (3) invariant, meaning that the characteristics of question items do not depend on the distribution of test taker ability parameters and the parameters that characterize test takers do not depend on the characteristics of question items (Hambleton, Swaminathan, & Rogers, 1991). This means that the results obtained will provide the same ability even though the questions are done by students who are less smart and smart or students from the lower middle class and students from the upper middle class will not give different student ability results.

The *Item Response Theory* (IRT) approach in this study was used to analyze and measure individual abilities in answering numeracy ability test questions. The IRT approach makes it possible to gain a deeper understanding of the level of numeracy ability of students. IRT makes it possible to measure numeracy ability more accurately than other traditional ones. The expected IRT model must have the following characteristics: (1) the characteristics of the question item do not depend on the group of test takers to whom the question item is subjected, (2) the score that describes the test taker's ability does not depend on *the test, (3) the model is expressed in the level of the* question item, not in the test level, (4) the level model does not require parallel tests to calculate the reliability coefficient, and (5) the model provides an accurate measure of each ability score (Hambleton, Swaminathan, & Rogers, 1991). The results of the IRT analysis can be used to design more targeted educational interventions. The IRT approach is able to identify students who need extra help and what types of assistance are most effective in improving numeracy skills. Based on the explanation above, it is necessary to analyze numeracy ability test items. This study has objectives, namely (1) to know the construct of numeracy ability test instruments for class VIII public junior high school students in Pekalongan Regency, (2) to know the quality of numeracy ability test instruments for class VIII public junior high school students in Pekalongan Regency, (3) to know the numeracy ability profile of class VIII public junior high school students in Pekalongan Regency.

Research Method

***Types of Research*.** This study used a quantitative approach by developing student numeracy ability test instruments. The purpose of this study was to see the characteristics of numeracy ability test items. The steps for developing test instruments adopted from (Istiyono, 2020) consist of eleven steps, namely 1) Determining the purpose of the test instrument, 2) Determining the competencies and materials to be tested, 3) Compiling the test item distribution matrix, 4) Compiling the test instrument *blueprint,* 5) Writing and designing test items, 6) Compiling scoring rubrics, 7) Validity of test instrument items, 8) Revision of test items, 9) Assembling instruments, 10) Carrying out test trials, and 11) Taking measurements with the main test instruments.

***Time and Place of Research*.** This research was carried out at the junior high school level within the scope of the education office of Pekalongan Regency, Central Java Province. Data sampling will be carried out from April to June 2023.

***Data, Instruments, and Data Collection Techniques*.** Data collection in this study was using test instruments. The test instrument is in the form of a student numeracy ability test done by students. This test has the purpose of determining numeracy ability. Test questions are questions with indicators contained in AKM.

***Data Analysis*.** Content validity is obtained by rational analysis of test content based on *expert judgment* (Allen & Yen, 1979). Content validity consists of advanced validity and logical validity (Allen &; Yen,

1979). Content validity has a *judgemental nature* where the results of the analysis are based on rational judgment from *experts*. Here we will see to what extent the agreement of experts can be proven empirically. Numeracy ability test instrument test data are analyzed to find evidence of construct validity, reliability, and item characteristics. Proof of construct validation is done with *Confirmatory Factor Analysis* (CFA) with the help of *R program software*. Construct validity is used to see or describe the extent to which instrument-measuring theories are used (Allen & Yen, 1979).

Reliability is an important aspect that can show the reliability of a measuring instrument. An instrument is said to be reliable if the instrument can be consistent in measuring a latent variable. High-reliability results will minimize the error rate in measurement (Retnawati, 2016). Reliability estimation can use construct reliability, where estimation can be done after proving construct validity through CFA when it has obtained a suitable model or *fit* model (Retnawati, 2016). Reliability after CFA analysis is classically reliability with *Alpha Cronbach*.

The trial of the student numeracy ability test instrument will produce student response data after working on the question items to be given. The data obtained will be analyzed using IRT modeling procedures for dichotomus and polytomus scores. The procedure carried out is testing IRT assumptions which then proceed to estimating question items and capabilities with a two-parameter logistical model.

Results and Discussion

Content Validity. *Expert judgment* provides a quantitative assessment of each item related to the instrument developed. *Expert judgement* gives a score of 1–5. The scores obtained are then analyzed using Aiken's V formula. Based on the validation process that has been carried out on the developed numeracy ability instrument, it can be seen in Table 1.

Table 1. Results of Numeracy Ability Content Validity Analysis

Items	V Aiken	Decision	Items	V Aiken	Decision
Men_1	0,7917	Valid	PGK_1	0,9167	Valid
Men_2	0,9167	Valid	PGK_2	0,8333	Valid
Men_3	0,8333	Valid	PGK_3	0,7917	Valid
Men_4	0,8333	Valid	PGK_4	0,8750	Valid
Men_5	0,8750	Valid	PGK_5	0,8333	Valid
PG_1	0,8333	Valid	PGK_6	0,9583	Valid
PG_2	0,7917	Valid	PGK_7	0,8333	Valid
PG_3	0,7917	Valid	PGK_8	0,8750	Valid
PG_4	0,7917	Valid	PGK_9	0,8333	Valid
PG_5	0,7917	Valid	PGK_10	0,8750	Valid
PG_6	0,9167	Valid	IS_1	0,8333	Valid
PG_7	0,8333	Valid	IS_2	0,8750	Valid
PG_8	0,9583	Valid	IS_3	0,8750	Valid
PG_9	0,8750	Valid	IS_4	0,7917	Valid
PG_10	0,9167	Valid	IS_5	0,8750	Valid
PG_11	0,9167	Valid	U_1	0,9583	Valid
PG_12	0,8750	Valid	U_2	0,9167	Valid
PG_13	0,8333	Valid	U_3	0,9167	Valid
PG_14	0,9583	Valid	U_4	0,9167	Valid
PG_15	0,9167	Valid	U_5	0,9167	Valid

Based on the content validation analysis with the Aiken V formula assisted by the R program can be seen in Appendix 3b and the category rating is 5 with a significance level of 0.029, the numeracy instrument is declared analytically valid if the Aiken V value. $\geq 0,79$ The results of the analysis can be seen in Table 10, it can be concluded that all items developed in a total of 40 items are classified as good with details of 5 matching items (Men), 15 multiple-choice items (PG), 10 complex multiple-choice items (PGK), 5 short-fill items (IS), and 5 description items (U). There were no missing items, but based on research considerations only 20 question items were used in conducting the trial. The selected items are items that represent aspects of numeracy,

namely algebra, numbers, geometry, and measurement, as well as data and uncertainty so for each aspect there are items that represent.

Construct Validity and Reliability. The final results of the construct validation analysis of numeracy ability were measured by 20 items, but of the 20 items there were items that were merged into one so that the CFA analysis for numeracy ability involved 12 items. The merged question items are question items that have similarities in terms of item indicators, namely 4 items for algebraic aspects, 2 items for number aspects, 2 items for geometry and measurement aspects, and four items for data and uncertainty aspects. Figure 1 presents a hypothetical visualization of a converse measurement model for numeracy ability.

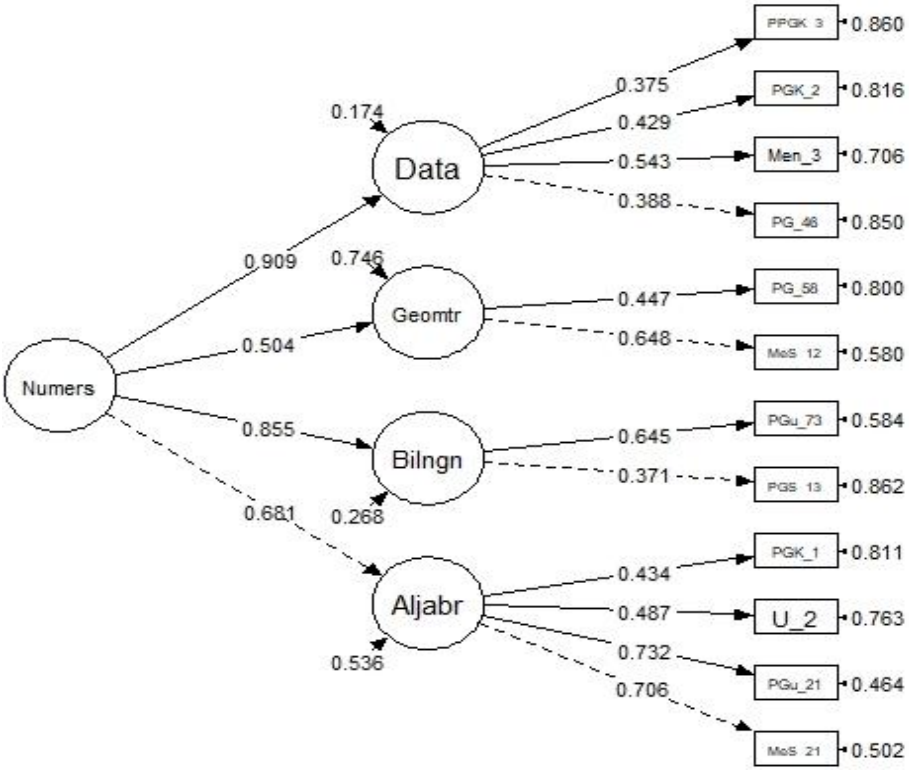

Figure 1. *Loading Factor Standardized Solution* Numeration

Based on empirical data, it will be seen how the model fits, the value of the loding factor, and the *t-value (*p-value). The compatibility of the numeracy ability measurement model is seen through the *chi-square (*p-value) value *of 0.05 and RMSEA (*Root Mean Square Error of Approximation) *< 0.08.* The value of the loding factor can be seen through a hypothetical model after CFA analysis as shown in Figure 1. Numeracy construct measurement model fit evaluation is a second-order CFA measurement model with four latent constructs, namely algebraic construct, number construct, geometric and measurement construct, and data construct and uncertainty. The fit of this measurement model is characterized by *chi-square* values = 64.766 with df = 50 *and p-value* = 0.078, and RMSEA values = 0.032. Based on the results of the evaluation of the suitability of the measurement model, it can be concluded that the empirical data obtained as a whole shows a match with the hypothetical model of measuring numeracy ability.

The load result *of the Loading Factor Standardized Solution first order for all items shows good convergent validity by considering* the Loading Factor Standardized Solution *and* the t-value *(*p-value). The results of the *Loading Factor Standardized Solution* analysis have been more than 0.3 and a significant factor load characterized by a *p-value* smaller than the maximum value p = 0.05 which has been set as a significance criterion. This means that they have good convergent validity, i.e. MeuS_21, PGu_21, U_2, and PGK_1 items for Algebraic constructs, PGuS_13 and PGu_73 items for Number constructs, MeuS_12 and PG_58 items for Geometry and Measurement constructs, and PG_46, Men_3, PGK_2, and PPGK_33 items for Data and Uncertainty constructs. The most dominant to less dominant aspects are Data and Uncertainty, Numbers, Algebra, and Geometry and Measurement. The dominant aspect can be seen from the results of the *second-order Loading Factor Standardized Solution*.

The next construct measurement model of numeracy ability is construct reliability using *Cronbach*'s Alpha in CFA to help measure the internal consistency of a set of measurable variables that contribute to latent factors. In simple terms*, Cronbach's Alpha value* is also considered a unidimensional index, which measures the extent to which a test measures a single factor (Gregory, 2000: 85). This means that reliability estimation using *Cronbach's Alpha is* able to provide information about the extent to which variables are measured in the CFA model. The latent variable of numeracy ability provides high construct reliability, namely the value of *Alpha Cronbach* (α) = 0.708.

Table 2. Eigenvalues Value Unidimensional Analysis

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	3,088	15,439	15,439
2	1,449	7,246	22,685
3	1,256	6,282	28,967
4	1,200	6,001	34,968
5	1,126	5,631	40,599
6	1,083	5,414	46,014
7	1,021	5,105	51,119
8	0,959	4,796	55,915
9	0,944	4,719	60,633
10	0,924	4,621	65,255
11	0,898	4,492	69,746
12	0,825	4,123	73,869
13	0,775	3,876	77,745
14	0,756	3,780	81,525
15	0,731	3,653	85,178
16	0,705	3,525	88,703
17	0,668	3,340	92,044
18	0,570	2,849	94,892
19	0,564	2,818	97,710
20	0,458	2,290	100,000

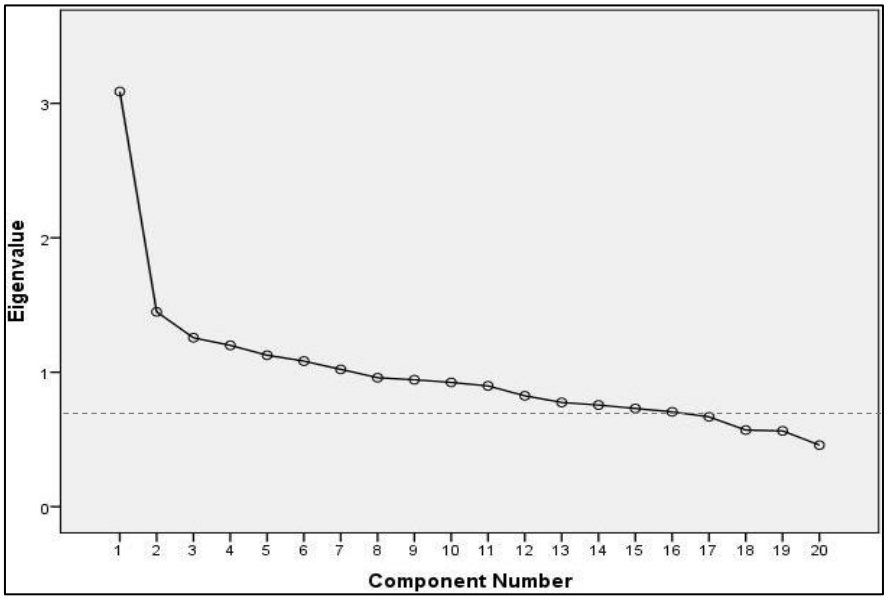

Figure 2. Unidimensional Scree Plot

Grain Characteristics. Unidimensional assumptions can be proven through construct validity using EFA which can be seen through the *Total Variance Explained Table, the* results of analysis using SPSS 23 are presented in Table 2 and *the Scree Plot* in Figure 2. Table 2 explains that there are 7 factors formed, but there

is one dominant factor, namely the first factor with *an eigenvalue* of more than 1, which is 3.088 as shown in Table 2. The number *of components* formed is 7 factors that can explain the variance of 51.119%. This explains that the numeracy ability instrument developed is able to distinguish student ability variances by 51.119%.

The proof of this unidimensional assumption is reinforced *by the scree plot* where it is proven that there is only one steep that is most dominant, namely the steep produced by the first factor. The *scree plot* graph looks from the first component/factor to the second component/factor steeply and from the second component/factor the graph is already sloping. This shows that there is one dominant factor, namely the numeracy ability of students. The resulting *screen plot* graph is shown in Figure 2.

Based on Figure 9, the chart shows a steep plot screen *chart pattern on* the first factor and starting to ramp up on the second factor, and so on. This proves that there is only one factor or factor that is dominant in the numeracy ability test instrument device is numeracy. The assumption of local independence can be fulfilled if the proof of unidimensional assumptions is also fulfilled (Mars, 2010 in Retnawati, 2014: 8). This means that if the unidimensional assumption is met, then automatically the assumption of local independence will also be fulfilled. This is because these two concepts are equivalent or equivalent (Lord, 1980; Lord & Novick, 1968 in Hambleton, Swaminathan, & Rogers., 1991: 11). The results of the unidimensional assumption test have been met, so automatically the assumption of local independence has also been fulfilled.

The third assumption test of *Item Response Theory* is the invariance of item parameters and capability parameters. Item parameter invariance means that item parameters will not affect/change if done by different groups of students while ability parameter invariance means that students' abilities will not be affected due to taking tests that have different levels of difficulty (Retnawati, 2014: 3). The assumption of invariance will be proven using the Rasch estimation model because it only considers the estimated difficulty of the item.

Proof of item invariance is carried out by dividing respondents into two groups, namely the male student group and the female student group. The next step is to estimate the difficulty of the item for the male student group and the item difficulty estimation for the female student group. Proof of the assumption of buitr invariance with the help of the program R generated *scatter plot* as in Figure 3.

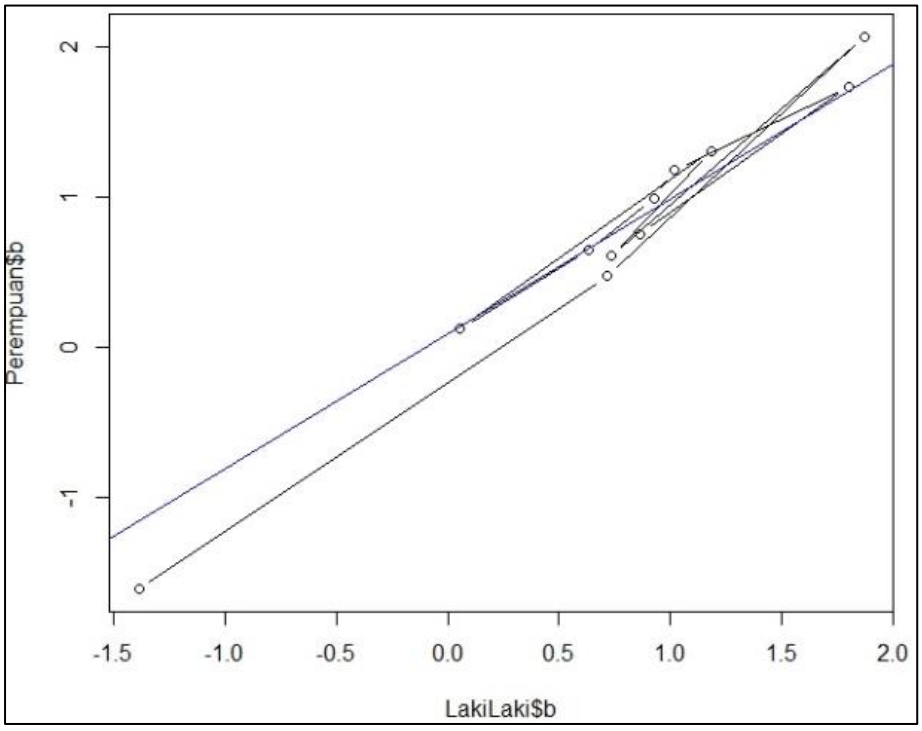

Figure 3. Item Parameter Invariance

Based on the *scatter plot* in Figure 3, it can be seen that there is no variation in the level of difficulty based on sex in the numeracy ability instrument. This proves that the level of difficulty estimated based on the group of men is almost the same as the level of difficulty estimated based on the group of women. This can be seen from the scores of male and female students are around a straight line, meaning that gender does not affect the variance of student scores in doing numeracy ability test instruments.

Analysis of item characteristics is carried out with the help of the R program. Dichotomus and polytomus data in the R program can be analyzed simultaneously. Dichotomus data were analyzed using the Rasch/1PL IRT model, while polytomus data were analyzed using the *Partial Credit Model* (PCM) model. This item characteristic analysis only looks at the difficulty level of the item so it uses a mixed model of the Rasch/1PL model and *the Partial Credit Model* (PCM) model.

The results of the analysis using the R program will produce item parameter estimates (difficulty level), model fit estimates (*fit model*), *Item Characteristic Curve* (ICC), *IFF,* SEM, and capability estimates. The data analyzed at this stage contained 20 items that had been proven valid in content and construct, namely 8 multiple-choice items, 3 short-fill items, 3 matchmaking items, 3 complex multiple-choice items, and 3 description items.

IRT analysis begins by looking at the *fit* model which means that the item is in accordance with the model used or the item can consistently function normally in making measurements according to the model used. The results of the model fit analysis are used to eliminate items that do not match the model used. Items that do not fit or do not fit indicate that there is a respondent's misconception of the item being tested. The fit model test *is* not suitable if it is based on *chi square values* because it will tend to reject if the sample is large (Hooper, Couglan, &; Mullen, 2008: 57). The *chi square* value is very sensitive to the sample size. The *chi square value will increase and lead to model rejection, if the number of samples is above 200 then* the chi square value will continue to rise so that there is a tendency to reject the null hypothesis (Haryono, 2016: 66). Model fit tests in addition to using *chi square* values can also be done by looking at the *infit value* to see the quality of items (Fisher, 2007: 1095).

The decision-making criterion for model fit using *infit is if* the infit value has a value ranging from 0.77 to 1.30 (Fisher, 2007: 1095). The study of infit decision criteria *is reinforced by (Keeves &; Alagumalai, 1999: 36) stating that the suitability of the model follows the rule that* the Item Characteristic Curve *(ICC) will be flat if* the *infit* value for the item is greater than 1.30 or less than 0.77. This, reinforced by (Adams & Khoo, 1996: 30) the item will be said to match or *fit* with the model used if the *infit value* ranges from 0.77 to 1.30. The results of the R-program-assisted model fit analysis can be seen in Table 3.

Table 3. Numeration Capability Model Fit Analysis Results

Items	Infit	Fit Model	Items	Infit	Fit Model
1	1,024	Fit	11	0,958	Fit
2	0,898	Fit	12	0,991	Fit
3	0,991	Fit	13	0,919	Fit
4	0,952	Fit	14	0,929	Fit
5	0,997	Fit	15	0,951	Fit
6	0,972	Fit	16	0,956	Fit
7	0,986	Fit	17	1,034	Fit
8	1,009	Fit	18	0,879	Fit
9	0,864	Fit	19	0,894	Fit
10	0,937	Fit	20	0,955	Fit

Based on Table 3, overall it can be seen from the *infit* value it can be concluded that the overall items used, namely 20 items analyzed using a mixture of Rasch and PCM models, it is proven that 20 items match or fit the model and can be used at the measurement stage.

The estimated parameter analyzed in this study is the difficulty level of the item (b). Test items can be tied well if the difficulty level of the item is in the interval -2.0 to 2.0 logit scale (Hambleton &; Swaminathan, 1985: 36). The characteristic curve of item 3 can be seen in Figure 4 of the items analyzed with a mixed model.

Dichotomous data is applied to the Rasch model with the parameter of the analyzed item is the difficulty level (b) and for the Rasch model the difference power parameter of all items is considered the same, namely 1 which affects all ICCs having the same slope. Analysis of item parameter estimation with the help of the R program for item parameter data from the analysis of the Rasch/1PL mixed model and the *Partial Credit Model* (PCM) model is presented in Table 4.

Based on Table 4, we can see the estimated parameters of the Rasch/1PL mixed model and the *Partial Credit Model* (PCM) model. The interpretation of the item characteristic curve for the dichotomus data in

Figure 12 is that the minimum probability required of the student answering correctly from an item is 50% or 0.5. Suppose that the PG_3 item for a dichotomous item such as Figure 4, the difficulty level can be estimated

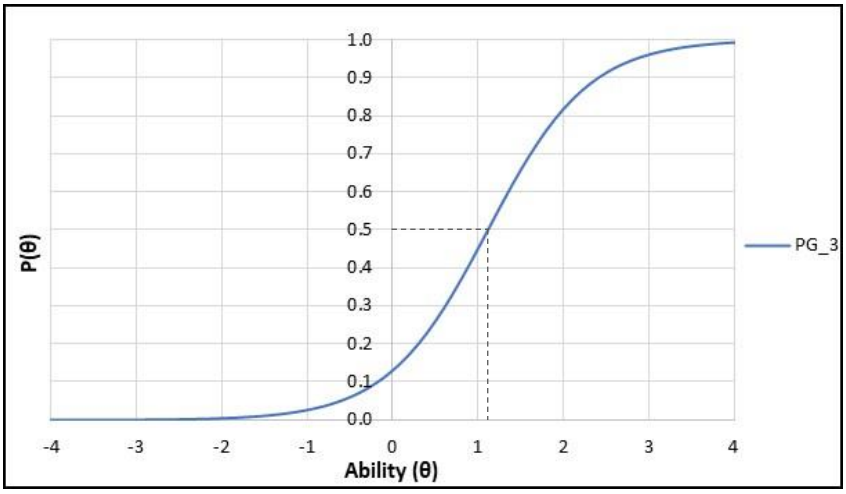

Figure 4. Item Characteristic Curve (ICC) Dichotomus Item PG_3

Table 4. Item Characteristic *Curve* (ICC) Numeration Capability

Items	***a***	***b***	b_1	b_2	**IRF**
PG_1	1,000	0,794	NA	NA	0,794
PG_2	1,000	1,762	NA	NA	1,762
PG_3	1,000	1,114	NA	NA	1,114
PG_4	1,000	0,962	NA	NA	0,962
PG_5	1,000	0,642	NA	NA	0,642
PG_6	1,000	0,100	NA	NA	0,100
PG_7	1,000	1,257	NA	NA	1,257
PG_8	1,000	0,658	NA	NA	0,658
IS_1	1,000	1,985	NA	NA	1,985
IS_2	1,000	0,565	NA	NA	0,565
IS_3	1,000	-1,519	NA	NA	-1,519
Men_1	1,000	NA	-2,147	1,685	-0,231
Men_2	1,000	NA	-1,360	0,531	-0,415
Men_3	1,000	NA	-1,291	1,328	0,018
PGK_1	1,000	NA	-0,626	-0,018	-0,322
PGK_2	1,000	NA	-0,589	0,811	0,111
PGK_3	1,000	NA	-1,505	0,737	-0,384
U_1	1,000	NA	-1,733	1,638	-0,047
U_2	1,000	NA	-0,254	0,139	-0,057
U_3	1,000	NA	-0,997	1,324	0,163

by drawing a straight line from P(θ) at a probability value of 0.5 to the right until it meets the curve, then drawn a straight line down which is the ability of (θ) 1.114. This means that it takes 1.114 abilities to correctly answer PG_3 item with a probability of 50%.

The interpretation of the item characteristic curve for the polytomus data in Figure 5 was analyzed using PCM which means that a high category score indicates greater ability than a lower score. This means that increasing the score requires certain abilities with a *threshold* or minimum ability or *step* parameter applies $b_1 < b_2 < b_3 < \ldots < b_n$. PCM analysis considers each category as a parameter step that students must pass to reach the correct answer. Polytomus curves will show the intersection between curves which are often referred to *as step* parameters and also as minimum opportunities and abilities (Saepuzaman *et al*, 2022:

273). Items with polytomus scores for numeracy instruments using 3 categories, namely 0, 1, and 2, will produce the step parameter $(b_1 < b_2)$.

Figure 5. Item Characteristic Curve (ICC) Polytomus Item Men_3

Table 4 shows that all polytomus items show b_1 and b_2 are increasing sequentially. The interpretation of the item characteristic curve for polytomus data will be given an example of Men_3 item as in Figure 5 showing the difficulty index of the step parameter shown through the intersection of the curve. For example, the intersection point between P1 and P2 represents the minimum opportunity and ability or can be said to be the first step parameter (δ) and the *intersection point between P2 and P3 represents the minimum opportunity and ability or can be said to be the* second step b_1 parameter (b_2). Figure 5 shows that the *step* 1 parameter is -1.291 and *the step* 2 parameter is 1.328. This means that students have a chance to score 1 if they have a minimum ability of -1.291, less than -1.296 then the greatest chance of getting a score of 0. In addition, the chance to get a score of 2 students must have a minimum ability of 1,328.

The item difficulty range for dichotomus data and polytomus data will be viewed through IRF to determine the difficulty level location index. IRF analysis is useful for looking at a single index for polytomus items (Ali, Chang, & Anderson, 2015: 2). Rasch/1PL mixed model and *Partial Credit Model (PCM*) model, so it is necessary to know the single index of difficulty level in PCM.

IRF has a more precise basis for representing the difficulty of the polytomus item as a whole. A single index using IRF also considers the item information function. Each test item has its own IRF value. The use of IRF can provide accurate information for each item. IRF can also be used to determine the difficulty level of an item, as well as IRF hypothetically has values ranging from $-\infty$ up to $+\infty$ (Gregory, 2000: 108). An item can be good if it has an item difficulty value ranging between -2 and +2 (Hambleton, Swaminathan, & Rogers, 1991). If viewed from the IRF results, it can be concluded that all items have a good level of difficulty.

The ability estimates of 599 students obtained from *the* R program output are presented on a logit scale between -4.0 to 4.0 Appendix 7e. The categorization of students' numeracy abilities is carried out using the normal distribution. The results of estimating student ability will be converted first into standard scores because they are closely related to follow-up actions in numeracy assessment. The logit scale resulting from the estimated ability to use MLE will first be converted into a scale of 0 to 100 for easy interpretation. The conversion results will be divided into four categories suggested by the Pusmenjar of the Ministry of Education and Culture, namely Proficient, Cakap, Basic, and Need Special Intervention (Ministry of Education and Culture, 2020: 29). The results of estimating students' numeracy ability after conversion can be seen in Table 5.

Table 5. Estimation of Student Numeracy Ability

Estimation Numeration Ability	Category	Sum	(%)
$X \geq \overline{x} + 1,5SB$	Skillful	36	6
$\overline{x} \leq X < \overline{x} + 1,5SB$	Clever	139	23
$\overline{x} - 1,5SB \leq X < \overline{x}$	Basis	390	65
$X \leq \overline{x} - 1,5SB$	Need Special Intervention	34	6
Sum		**599**	**100**

Based on Table 5, it shows that the numeracy ability of junior high school students in Pekalongan Regency ranges from 26.1605 to 77.5433 with an ideal average of 51.8519. The numeracy ability of junior high school students in Pekalongan Regency shows that there are 36 students out of 599 students classified as proficient with a percentage of 6%, 139 students out of 599 students classified as proficient with a percentage of 23%, 390 students out of 599 students classified as basic with a percentage of 65%, and 34 students out of 599 students classified as needing special intervention with a percentage of 6%.

The validity and reliability of instruments based on the IRT approach can be seen through the value of the information function and the *Standard Error of Measurement* (SEM). The function of measurement result information is very important in suckling instruments. The information obtained can be used to select items or can provide information regarding the strength of the item in measuring latent capabilities to be used according to its measurement purpose. In addition, through the results of the information function can also be known *Standard Error of Measurement* (SEM) or errors in measurement. The SEM value can be obtained from the inverse square root of the information function. The curves of the information function and SEM will intersect each other as shown in Figure 6.

Figure 6. Information Function Curves and SEM

Based on Figure 6, the maximum value of the numeracy ability test information function is 10.503 (θ) at 0.0 and SEM at 0.308. A good or reliable test instrument when the *Total Information Function* (TIF) value is $\geq$ 10 according to Hambleton (in Wiberg, 2004). This means that the instruments used include good and reliable instruments to measure students' numeracy abilities based on the results of the R program analysis. In addition, there is an intersection of information function curves and SEM intersecting at θ = -2.4 and 3.2 which means that the instrument is very suitable or will provide accurate information or greater than Standard *Error Measurement* (SEM) from the left intersection to the right intersection or it can be said that the numeracy ability instrument will be reliable and provide accurate information if given to students who have the ability =

θ -2.4 to = θ 3.2. This range shows that the numeracy ability instrument is able to measure students' abilities with a fairly wide range.

Conclusion

Based on the results of the analysis, it can be concluded that the construction of numeracy ability instruments for grade VIII State Junior High School students, is related to the content of algebra, numbers, geometry, and measurement, as well as data and uncertainty. In addition, it uses personal, socio-cultural, and scientific contexts, using cognitive levels of understanding, application, and reasoning. The quality of the numeracy ability instrument is declared valid and reliable, and the construct validity of all items is fit as seen from the *Loading Factor Standardized Solution* value of more than 0.3 and *p-value* < 0.05 as well as high category reliability and estimated item characteristics show that the question items are included in the category both in terms of difficulty. The numeracy ability of junior high school students in Pekalongan Regency shows that there are 36 students out of 599 students classified as proficient with a percentage of 6%, 139 students out of 599 students classified as proficient with a percentage of 23%, 390 students out of 599 students classified as basic with a percentage of 65%, and 34 students out of 599 students classified as needing special intervention with a percentage of 6%.

Cite this article as: Alan Rifqi Kamal, Edi Istiyono (2023). Analysis of numeracy ability test item characteristics grade VIII students with mixed model item response theory (IRT) approach. *Challenges of Science*. Issue VI, 2023, pp. 184-195. https://doi.org/10.31643/2023.22

References

Abikak, Y., Kenzhaliyev, B., Retnawati, H., Gladyshev, S., & Akcil, A. (2023). Mathematical modeling of sulfuric acid leaching of pyrite cinders after preliminary chemical activation. Kompleksnoe Ispolzovanie Mineralnogo Syra = Complex Use of Mineral Resources, 325(2), 5–13. https://doi.org/10.31643/2023/6445.12

Adams, R. J., & Khoo, S. T. (1996). Quest: The Interactive Test Analysis System Version 2.1. The Australian Council for Education Research.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psycholodical Measurement, 45 (1), 131-142. https://doi.org/10.1177/0013164485451012

Ali, U. S., Chang, H. H., & Anderson, C. J. (2015). Location Indeces for Ordinal Politomus Items Based on Item Response Theory. Princeton, NJ: Educational Testing Service, 1-13. http://dx.doi.org/10.1002/ets2.12065

Allen, M. J., &; Yen, W. M. (1979). Introduction to Measurment Theory. Brooks/Cole.

Fisher, W. P. Jr. (2007). Rating scale instrument quality criteria. Rasch Measurement Transaction, 21(1), p.1095. https://www.rasch.org/rmt/rmt211m.htm

Gregory, R. J. (2000). Psychological Testing: History, Principles, and Applications. United States of America: Allyn & Bacon, Inc.

Gunawan, M. A., Yunus, M., A'yun, Q., &; Yuniarti, I. A. (2020). Practical Guide to Analysis of Assessment Instruments Using Item Response Theory (IRT). Yogyakarta: Nuta Media

Hambleton, R. K., &; Swaminathan, H. (1985). Item Response Theory: Principles and Application. New York: Kluwer-Nijhoff

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2). Newbury Park, CA: Sage Publications Inc.

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.

Istiyono, E. (2020). Development of Assessment Instruments and Analysis of Physics Learning Outcomes with Classical and Modern Test Theory (second edition). Yogyakarta: UNY Press.

Ministry of Education and Culture (2020). AKM Problem Development Design. Jakarta: Directorate General of Early Childhood, Basic Education, and Secondary Education of the Ministry of Education and Culture.

Mardapi, D. (2017). Measurement, Assessment, and evaluation of Education (2nd). Yogyakarta: Parama Publishing

Megawati, L.A., Sutarto, H. (2021). Analysis numeracy literacy skills in terms of standardized math problem on a minimum competency assessment. Unnes Journal of Mathematics Education, 10(2), 155-165. DOI: 10.15294/ujme.v10i2.49540

OECD (2019), PISA 2018 Assessment and Analytical Framework, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b25efab8-en.

Puspaningtyas, N. D., &; Ulfa, M. (2020). Training on numeracy literacy-based math problems for high school students it fitrah insani. Journal of Mathematics and Natural Sciences Community Service and Mathematics and Natural Sciences Education, 4(2), 137-140.

Retnawati, H. (2014). Item response theory and its application. Yogyakarta: Parama Publishing.

Retnawati, Heri. (2016). Quantitative Analysis of Research Instruments. Yogyakarta: Parama Publishing.

Saepuzaman, D., Istiyono, E., &; Haryanto, H. (2022). Characteristics of fundamental physics higher-order thinking skills test using item response theory analysis. Pagem Journal of Education and Instruction, 12(4), 269-279. https://doi.org/10.47750/pegegog.12.04.28

Wright, B., & Stone, M. (1999). Measurement Essential (2nd ed). Wilmington: Wide Range, INC.