International Journal of Teaching and Learning (INJOTEL) e-ISSN: 3025-8308
Vol. 2 No. 12, December 2024, pages: 3376~3391

DEVELOVING QUANTITATIVE LITERACY ASSESSMENT TOOLS FOR INDONESIAN SECONDARY SCHOOL STUDENTS

Khaerun Nisa*
Pascasarjana Universitas Negeri Jakarta, Indonesia
E-mail: khaerun.nisa@mhs.unj.ac.id

Siska Merrydian
Pascasarjana Universitas Negeri Jakarta, Indonesia
E-mail: siska.merrydian@mhs.unj.ac.id

Awaluddin Tjalla
Pascasarjana Universitas Negeri Jakarta, Indonesia
E-mail: awaluddin.tjalla@unj.ac.id

Iva Sarifah
Pascasarjana Universitas Negeri Jakarta, Indonesia
E-mail: ivasarifah@unj.ac.id

Abstract

Quantitative literacy ability is an important skill in the digital era 4.0. This research aims to develop a quantitative literacy instrument for class VII students that is valid, reliable, free from Differential Item Functioning (DIF), and capable of measuring abilities accurately. The development method uses a 4D model (Define, Design, Develop, Disseminate), limited to the Develop stage. The instrument consists of 36 questions based on six dimensions of quantitative literacy (interpretation, representation, calculation, assumptions, analysis, communication) which are validated by experts with scores CVI 0.94. Trials were carried out on 220 class VII students, and analysis using the Rasch Model showed high reliability (item reliability 0.99; person reliability 0.85) and good precision (standard error 0.25 logit). The distribution of difficulty levels of questions is in line with average to medium student abilities, although the coverage of extreme abilities needs to be improved. The instrument is considered valid, reliable and precise, but it is recommended to add questions with a higher level of difficulty and review questions that are too easy or difficult to make them more representative.

Keywords: Quantitative Literacy, Secondary School

INTRODUCTION

Quantitative literacy (QL) skills are one of the essential competencies in facing the challenges of the 21st century, especially in supporting data-based decision making in various aspects of life. Quantitative literacy not only includes calculation skills, but also the ability to understand, analyze, and apply mathematical concepts in real situations (Steen, 2001). This competency is very relevant in the context of globalization and digitalization, which requires individuals to be able to manage data-based information well.

In Indonesia, students' quantitative literacy skills are still a serious challenge. Based on the results of the 2022 Program for International Student Assessment (PISA), the average mathematics score of Indonesian students was recorded at only 366, far below the OECD average of 472. In addition, only 18% of Indonesian students reached the minimum proficiency level (Level 2), while in OECD countries, more than 69% of students are at that level (OECD, 2022). This shows that many Indonesian students have difficulty applying mathematical concepts to real life situations. One factor contributing to low quantitative literacy is the limited availability of assessment tools relevant to local contexts, which tend to only measure procedural aspects of mathematics and do not cover the analytical skills needed in real-world situations (Steen, 2001; OECD, 2022) indicating large gaps in numeracy skills needed to face the challenges of the modern world. This condition is a serious concern because quantitative literacy is an important foundation in various scientific disciplines and work skills.

In line with this, the Indonesian government has attempted to increase numeracy literacy through policies such as the Minimum Competency Assessment (AKM), which is part of the national assessment. AKM aims to measure students' basic abilities in literacy and numeracy as indicators of educational quality (Kemendikbud, 2020). However, the AKM instrument still focuses on measuring numeracy in general without paying special attention to more complex aspects of quantitative literacy, such as the ability to analyze contextual data or solve problems based on real situations. In addition, although policies such as the Minimum Competency Assessment (AKM) have been implemented to assess students' literacy and numeracy competencies, the assessment instruments used are not sufficient to provide a comprehensive picture of Indonesian students' quantitative literacy abilities. Existing assessment instruments still largely ignore the applicable aspects of quantitative literacy, such as the ability to understand statistical data or use mathematics to solve complex problems in real life contexts (Isnawati, Rahayu, & Pratiwi, 2021). In addition, these instruments are generally generic and do not take into account the local and cultural characteristics of Indonesian students, which has an impact on the lack of relevance of assessment tools to students' daily experiences (Kemendikbud, 2020). This results in students being less skilled in connecting mathematical knowledge with real-world challenges, such as graph interpretation, budget planning, or risk analysis.

This research aims to develop a quantitative literacy assessment tool that is contextual and relevant for secondary school students in Indonesia. This tool is designed to not only measure students' abilities in mathematical aspects but also evaluate their skills in using quantitative data critically in various practical situations. This approach is expected to be able to provide a more comprehensive picture of students' quantitative literacy levels as well as help educators identify weaknesses that need to be corrected. The development of this assessment tool has strategic value in supporting improving the quality of education in Indonesia. By using instruments that are evidence-based and relevant to local needs, the results of this research are expected to make a significant

contribution to improving competency-based learning and achieving sustainable national education development targets.

RESEARCH METHOD

This type of research is quantitative research involving Thiagarajan's (1974) Research and Development model, namely 4D (Define, Design, Develop, Disseminate). in the procedure for developing this quantitative literacy instrument, it is limited to the 3D stage (Define, Design, Develop) described as follows:

Define Stage

a. Conceptual Definition

Quantitative literacy is the ability to reason to solve problems in the form of numbers, arithmetic and statistics from various contexts and everyday life.

b. Operational Definition

Quantitative literacy is the ability to reason in managing information in the form of numbers and statistics from various contexts and daily life as measured through indicators, 1) interpretation ability, 2) representation ability, 3) calculation ability, 4) analysis ability, 5) application ability/ analysis, and 6) communication skills.

Design Stage

The Design Stage is a test design process that includes:

a. Grid Arrangement

Mardapi (2008) states that the steps for compiling a grid include: 1) writing a general objective, 2) making a list of discussion points, 3) determining indicators, 4) determining the number of questions. Before creating a grid, there are several things that must be considered, namely the number of questions that will be designed. In Neil's (2011) opinion, one domain contains at least 3 items. Then Suminto and Widhiarso (2015) stated that the number of items created by researchers must be two or three times more than the target number of items, with the reason that if there are items that do not pass the selection, there are still remaining items in reserve. Based on this theory, the number of items targeted to pass item selection is 18 items, because there are 6 measurement domains. Then the number of items that must be made is at least 2 times more than the target item, namely 36 items. The following is presented in Table 1 a grid of quantitative literacy instruments.

Table 1. Quantitative Literacy Instrument Grid

No.	Dimensions		Question Indicator	No. Question	Number of Questions
1.	Interpretation	1.	Examining information in graphical form in solving problems	1, 2, 3,	3

		2.	Examining tabular information in solving problems	4, 5, 6,	3
2.	Representation	1.	Conceptualize information into geometric pattern images	7, 8, 9,	3
		2.	Conceptualize information into the form of a mathematical model	10, 11, 12,	3
		3.	Conceptualize information in the form of diagrams, graphs and tables	13, 14, 15	3
3.	Calculation	1.	Using addition and subtraction arithmetic operations in solving mathematical problems	16, 17, 18,	3
		2.	Using multiplication and division arithmetic operations in solving mathematical problems	19, 20, 21,	3
		3.	Using mixed arithmetic operations in solving mathematical problems	22, 23, 24,	3
4.	Analysis	Analyzing information in the form of story questions in solving problems		25, 26, 27, 28,	4
5.	Assumption	Interpreting the results of problem solving from information in the form of story questions		29, 30, 31, 32,	4
6.	Communication	Conceptualize problem solving in the form of mathematical models		33, 34, 35, 36	4
Number of Questions					**36 Questions**

b. Determining Test Form and Test Length

The form of test used in this research is a multiple choice test, with the reason that it covers a wide range of material and takes a short time to complete it. Then the length of the test is shown in the following table:

Table 2. Test Length

No	Dimensions	Number of Questions	Time Estimate (Nitko, 1996)	Total time

1.	Interpretation	6	40-60 seconds/question	240-360 seconds/question
2.	Representation	9	70-90 seconds/question	630-810 seconds/question
3.	Calculation	9	2-5 minutes/question	18-45 minutes/question
4.	Assumption	4	70-90 seconds/question	280-360 seconds/question
5.	Analysis	4	2-5 minutes/question	8-20 minutes/question
6.	Communication	4	2-5 minutes/question	8-20 minutes/question
		36 items		**53,10 -110,30 minutes**

c. Item Writing

Writing quantitative literacy items in this research takes into account several aspects: 1) the relevance of the item to the quantitative literacy dimension, 2) the relevance of the item to the question indicators, 3) clarity of the main question, 4) the logicality of all answer options, 5) the standardness of the language used, and 6) the functionality of case/discourse descriptions, pictures, graphs and tables.

Development Stage

The Develop stage is a test development process based on a previous design which includes:

a. Expert Assessment

The expert assessment aims to determine the validity of the instrument's content. Content validity is the accuracy between the content of the test and the construct to be measured. Goodwin and Leech (2003) stated that validity based on test content is based on logical analysis and expert evaluation of measurement content such as item points, item formats, and the sentences that make up them. This opinion is also reinforced by Mardapi (2008) who states that content validity is related to the extent to which the item covers all the material or materials to be measured, which is analyzed through the assessment of several experts. The number of experts used to assess this development product is 3 experts consisting of 2 mathematicians and 1 measurement expert. This refers to the opinion of Lynn (1986) which states that the number of experts used in expert validation is a minimum of 3 experts and no more than 10 experts. The expert assessment data analysis technique used is the Content Validity Index/CVI. This technique is a proportion of expert assessments of content based on items that get a score of 3 or 4 (Polit and Beck, 2006). Product development criteria can be said to be content valid if the CVI proportion is > 0.60. This refers to the opinion of Rempusheski and O'Hara (2005) who stated that the recommended CVI proportion ranges from 0.60 to 1.0.

b. Trials

Field trials are carried out to determine empirical information about an item, such as the validity of the internal structure, reliability of the item and DIF detection instruments and level of difficulty. Apart from that, through trials you can also find out the person's reliability and ability. The trial of the quantitative literacy instrument in this research was carried out on class VII students at SMP Lab School Jakarta with a total of 220 students.

c. Item Analysis

Item analysis aims to determine the characteristics of quantitative literacy instrument items, such as internal validity, item and instrument reliability, DIF detection, and item difficulty level. The characteristics of the items in this research use the Rasch Model approach, guided by the following table:

Table 3. Item Analysis Guidelines

No	Analysis Aspect	Guidelines
1	Internal Validity	Valid internally if outfit ZSTD value: -2,0 < ZSTD < +2,0 (Boone, dkk, 2014).
2	Item reliability	Item reliabel jika nilai Item Reliability ≥ 0,67 (Fisher, 2007)
3	Reliability Instrument	The instrument is reliable if the coefficient reliability KR-20 ≥ 0,70 (Naga, 2013)
4	DIF Detection	Items are DIF free if probability Mantel-Haenszel > 0.05
5	Difficulty Level	Logit scale on Item Measure

d. Person Analysis

Person analysis aims to determine person characteristics such as reliability and ability. Person characteristics in this study were analyzed using the Rasch Model approach, guided by table 3.4 below:

Table 4. Person Analysis Guidelines

No	Analysis Aspect	Guidelines
1	Person Reliability	Person is reliable if the value of Person Reliability ≥ 0,67 (Fisher, 2007)
2	Difficulty Level	Logit scale on Person Measure

e. Assembling Tests

Assembling a test is the activity of arranging items into a single test unit (Mardapi, 2008). Things that must be considered when assembling the test include question validity, reliability, and undetectable DIF.

Data collection technique

The data collection techniques used in this research are questionnaire techniques and test techniques. Sequentially, these techniques are used to collect expert assessment data on the instruments being developed, and to collect empirical data on quantitative

literacy.

Data Analysis Techniques

The data analysis techniques used in the research are Content Validity Index (CVI) and Rasch Model. CVI is used to determine the content validity of the instrument being developed, while the Rasch Model is used to determine item and person characteristics empirically. Briefly, the data analysis above is presented in table 5 below.

Table 5. Data Analysis Techniques

No	Aspect Analysis	Approach Analysis	Criteria
1	**Content Validity**	Conten Validity Index (CVI)	Content validity if the CVI proportion is > 0.60 Rempusheski and O'Hara, 2005)
2	**Rasch Model Assumptions**		
	Item and Person Fit	Rasch Model	Item and Person fit if -2,0 < ZSTD < +2,0 (Boone, dkk, 2014)
	Unidimensional	Rasch Model	the unidimensional assumption is met if test variant value > 20 % (Reckase, 1979
	Independence Local	Rasch Model	The local independence assumption is met if the residual correlation value < 0,20 (Christensen, dkk, 2016).
	Invariance Group	Rasch Model	Group invariance checks can be guided by improvements Pure score along with level ability (Kang, & dkk, 2018)
3	**Item Characteristics**		
	Validity Internal	Rasch Model	Valid internally if the ZSTD outfit value is: -2,0 < ZSTD < +2,0 (Boone, dkk, 2014)
	Reliability Items and Instrument	Rasch Model	The instrument is reliable if the coefficient reliability KR-20 $\geq$ 0,70 (Naga, 2013) An item is reliable if the Item value Reliability $\geq$ 0,67 (Fisher, 2007)
	DIF Detection	Rasch Model	Items are DIF free if probability Welch $\geq$ 0,05
	Level Difficulty	Rasch Model	Logit scale on person Measure.
4	**Person Characteristics**		
	Ability	Rasch Model	Logit scale on person Measure.

Person		
Reliability Person	Rasch Model	Person is reliable if the value of Person Reliability ≥ 0,67 (Fisher, 2007)

RESULTS AND DISCUSSION

Expert Validation

In this research, 36 items were developed and involved 3 experts (2 mathematicians and 1 measurement expert) to test content validity. By using the Content Validity Index (CVI), it was found that each item was assessed based on the proportion of CVI from the three experts, and the average CVI value (Mean CVI) was calculated to determine its validity. Overall, the Mean CVI value shows very good results, with an overall average of 0.94. Of the total 36 items assessed, 26 items obtained the highest Mean CVI of 1.00, which reflects full agreement among experts regarding the validity of the items. The item with the lowest Mean CVI is item 13, with a score of 0.61, but it is still declared valid. All "Valid" items indicate that they meet the necessary validity criteria and indicate fairly good overall item quality, based on expert evaluation.

Dichotomous Rasch Model Analysis Assumption Test

Unidimensional

```
TABLE 23.0 Data.xls                                  OUTPUT  Nov 19 2024 11:29
INPUT: 220 PERSON  36 ITEM  REPORTED: 220 PERSON  36 ITEM  2 CATS WINSTEPS 5.3.1.0
-------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                    Eigenvalue   Observed   Expected
Total raw variance in observations     =     58.4437 100.0%         100.0%
  Raw variance explained by measures   =     22.4437  38.4%          38.0%
    Raw variance explained by persons  =      9.9935  17.1%          16.9%
    Raw Variance explained by items    =     12.4502  21.3%          21.1%
  Raw unexplained variance (total)     =     36.0000  61.6% 100.0%   62.0%
    Unexplned variance in 1st contrast =      3.4836   6.0%   9.7%
    Unexplned variance in 2nd contrast =      1.8227   3.1%   5.1%
    Unexplned variance in 3rd contrast =      1.7610   3.0%   4.9%
    Unexplned variance in 4th contrast =      1.6605   2.8%   4.6%
    Unexplned variance in 5th contrast =      1.5692   2.7%   4.4%

        STANDARDIZED RESIDUAL VARIANCE SCREE PLOT
```

Unidimensionality is seen in the raw variance explained by measure located in the observed column. The unidimensional assumption is met if the amount of raw variance explained by measure is > 20% (Reckase, 1979). It can be seen that the value is 38.4% > 20%. Further dimensional analysis can be proven through the eigenvalue units column (Huberty et al., 2013; Kaliski et al, 2013), the values obtained sequentially, namely: 3.4836; 1.8227; 1.7610; 1.6605; 1.5692. The variance that cannot be explained sequentially is: 6.0%, 3.1%, 3.0%, 2.8% and 2.7%. The variance values are in the 2-6% category. Thus, empirically the instrument is unidimensional and builds construct validity. It can be concluded that the unidimensional requirements are met.

Local Independence

The local independence criterion is violated if the residual correlation between pairs of items is positive and > 0.30 (Aryadoust et al., 2020). In the table below, there are no pairs

of items that have a residual correlation in the positive direction and greater than 0.30. Therefore, it can be concluded that the local independence assumption is met.

TABLE 23.99 Data.xls OUTPUT Nov 19 2024 11:29
INPUT: 220 PERSON 36 ITEM REPORTED: 220 PERSON 36 ITEM 2 CATS WINSTEPS 5.3.1.0

LARGEST STANDARDIZED RESIDUAL CORRELATIONS
USED TO IDENTIFY DEPENDENT ITEM

CORRELATION	ENTRY NUMBER	ITE	ENTRY NUMBER	ITE
.30	26	B26	27	B27
.30	27	B27	32	B32
.29	27	B27	29	B29
.28	27	B27	30	B30
.25	20	B20	21	B21
.25	4	B4	5	B5
.23	24	B24	27	B27
.23	27	B27	31	B31
.22	3	B3	14	B14
.22	30	B30	32	B32
.22	4	B4	9	B9
-.32	5	B5	27	B27
-.28	5	B5	29	B29
-.27	15	B15	27	B27
-.27	15	B15	30	B30
-.24	15	B15	29	B29
-.23	5	B5	31	B31
-.23	5	B5	32	B32
-.22	13	B13	22	B22
-.22	5	B5	26	B26

Monotonization

The nature of monotonization (Andrich, 2011) explains that positive sequential threshold distances are not isolated and it is said that response categories can be interpreted as an ordinal scale. Analysis shows that there is an increase in the value in the Observed Average column from negative to positive. The nature of monotonication can be seen from the observed average column where the values must tend to increase monotonically. Based on the table below, it can be seen that the observed average value tends to increase from -1.06 to 1.43. So it can be concluded that the monotonization assumption is met.

TABLE 3.2 Data.xls OUTPUT Nov 19 2024 11:29
INPUT: 220 PERSON 36 ITEM REPORTED: 220 PERSON 36 ITEM 2 CATS WINSTEPS 5.3.1.0

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	COHERENCE M->C	COHERENCE C->M	RMSR	ESTIM DISCR	
0	0	3399	43	-1.06	-1.06	.99	1.03	75%	70%	.4369		0
1	1	4521	57	1.43	1.43	1.00	1.11	79%	83%	.3476	1.00	1

OBSERVED AVERAGE is mean of measures in category. It is not a parameter estimate.
M->C = Does Measure imply Category?
C->M = Does Category imply Measure?

If the assumptions or prerequisites for the Rasch Dichotomy Model have been tested, then data analysis can be carried out using the Rasch Model.

Person Fit

The person is fit or not for the model based on outfit statistics, the value used is Outfit ZSTD, if the value is > 1.96 it indicates that the person is not fit for the model. Based

on the winstep output in table 6.1, 8 people are found who do not fit the model, namely 67, 179, 216, 33, 181, 147, 111, and 14, so these people can be discarded.

Item Fit

Based on the item fit analysis, it can be determined how many good items meet the Rasch Model criteria. To determine the quality of an item empirically, William P. Fisher's provisions can be used:

a) Item Model Fit Mean-Square Range Extremes or MNSQ outfit value 0.5 – 2.0
b) Outfit Z-Standard Value (ZSTD): -2.0 < ZSTD < +2.0
c) Outfit Measure Correlation (Pt Mean Corr) value 0.32 logit < Pt Measure Corr < 0.8 logit

Question items are said to be unfit and must be replaced if they do not meet the three criteria above, but question items are still said to be fit or retained if they meet at least the two criteria above (Sumintono & Widhiarso, 2015). The Jmle Measure column is item level information on a logit scale. In the table above, the items are ordered from most difficult to easiest. The most difficult item is Item 2 with a difficulty level of 6.46 logits and the easiest item is Item 9 with a difficulty level of -3.38 logits. Column Model S.E. shows the standard error for each item which is a statistic that describes how reliable the estimation results of the item parameters are in representing the population. And the Infit column is reported in two forms, namely mean-square form (MNSQ) and z statistical form (ZSTD). In making a decision whether an item fits the model or not, the value used is Infit MNSQ. The Outfit column is also reported in two forms, namely the mean-square form (MNSQ) and the z statistical form (ZSTD). In making a decision whether an item fits the model or not based on Outfit statistics, the value used is Outfit MNSQ. The MNSQ Outfit Criteria 0.5-1.5 shows that items 4 (2.24) and 11 (2.19) do not fit the model. So the item is discarded.

Item Measure

Based on the Winsteps output in table 13.1, it can be seen that the difficulty level of the items has been sorted from highest to lowest difficulty level. The most difficult item is item 2, while the easiest item is item number 9. A high measure value indicates that the item has a high level of difficulty. This correlates with the total score, where a small number of correct answers in the total score correlates with a higher measure value. The number of correct answers to a question can be seen in the Total Score, while the number of answerers to a question can be seen in the Total Count.

Based on guidelines from Suminto and Widhiarso (2015), the measure value resulting from the analysis shows the level of difficulty of each item, which is classified into four categories: very easy, easy, difficult, and very difficult, described as follows:

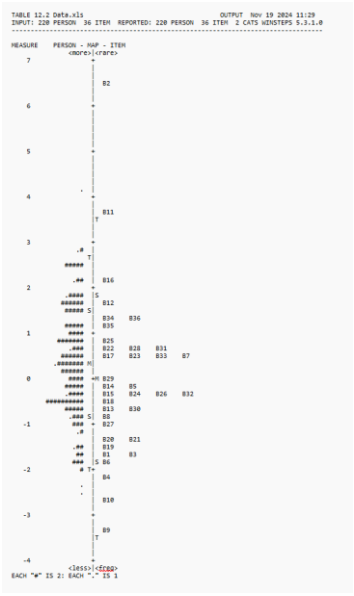
1. **Very Easy Item (Measure Value < -1)**
   Items included in this category have a very low level of difficulty, so they are easy for most respondents to answer correctly. These items include the numbers: **20, 21, 19, 1, 3, 6, 4, 10, and 9.** These items may not be able to differentiate the abilities or characteristics

of respondents because almost all respondents tend to be able to answer them correctly. These items need to be considered, whether they are relevant or too easy.

2. **Easy Items (Measure Value -1 to 0)**
   Items in this category are still relatively easy, but slightly more challenging than the first category. The items included are the numbers: **5, 14, 32, 26, 15, 24, 18, 30, 13, 8, and 27.** These items are good enough to measure respondents with lower ability, but may still be less sensitive in differentiating ability. higher respondents.
3. **Difficult Items (Measure Value 0 to 1)**
   This category includes items that have a higher level of difficulty, which can only be answered correctly by respondents with better abilities. These items include the numbers: 25, 31, 22, 28, 33, 23, 7, 17, and 29. Items in this category are good for identifying respondents with moderate to high levels of ability. However, the distribution of difficulty needs to be balanced so that all ability levels are represented.
4. **Very Difficult Item (Measure Value > 1)**
   Items in this category have a very high level of difficulty, so only a few respondents were able to answer them correctly. These items include the numbers: 2, 11, 16, 12, 36, 34, and 35. These very difficult items can be useful for measuring respondents with the highest abilities, but too many items in this category can cause reliability problems or bias against respondents with the highest abilities. low.

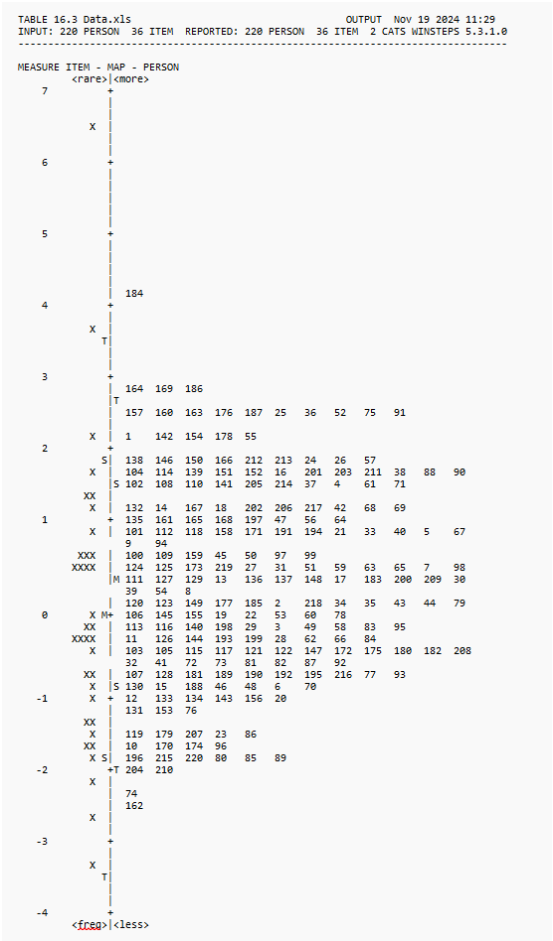
A balanced distribution of item difficulty levels is essential in building reliable and valid measuring tools. Based on these results, it was found that the majority of items were in the easy or difficult category, indicating that the scale had quite varied levels of difficulty. However, there are quite a lot of items in the very easy and very difficult categories. This can indicate potential bias in the measurement tool, where some items are too extreme to represent the entire population. Further evaluation of the distribution of items in each category is necessary to ensure this scale covers respondents' abilities as a whole, from low to high level. Adjustments to items that are too easy or too difficult may be necessary to increase the sensitivity of the measuring instrument in identifying differences in abilities or characteristics of respondents.

Wright Map Item



The Wright Map Item above provides an overview of the distribution of respondents' abilities and the level of difficulty of items in one instrument. In general, the majority of respondents had abilities that gathered around the measure value 0, reflecting average abilities, while a small number of respondents showed very high (>3) or very low (<-3) abilities. Most items also fall in the -1 to 1 range, indicating a level of difficulty that is in line with the respondent's average ability. However, there are several items that stand out as outliers, such as B2 with a very high level of difficulty (measure = 7), which may only be answered by respondents with very high ability, as well as items such as B4, B9, and B10 which are very easy (measure < -3), so it may be less informative in distinguishing respondents with low ability. In addition, there is a gap in the range of difficulty levels between measures 3 to 6, which potentially results in a lack of scope for measuring respondents with abilities at that level. Therefore, it is necessary to evaluate items that are too extreme to ensure their relevance and suitability for the target population. The addition of items of intermediate difficulty can help cover a wider range of abilities and improve the overall quality of the instrument.

Wright Map Person



The person Wright Map above describes the distribution of individual abilities (person abilities) and item difficulty levels (item difficulty) on the same logit scale. In the map above, individual abilities are spread from +7 to -4 logits, with the majority hovering around 0 to +3 logits. On the other hand, item difficulty ranges from around +4 to -4 logits, with the majority of items being in the 0 to -1 logits. This shows that most individuals have abilities that match the difficulty level of the dominant item. However, there is an imbalance in the extreme region, individuals with high ability (logit +4 to +7) do not have items that are difficult enough to measure their ability, while some items at logit -3 to -4 appear to be too difficult for most individuals because only few participants were in this range. Therefore, this test tends to be effective for measuring the abilities of individuals at average to medium levels of ability, but is less than optimal for measuring individuals with extreme abilities, both very high and very low. Improvements could be made by adding more difficult items to accommodate high-ability individuals and considering revising or deleting items that are too difficult to improve balance and measurement coverage.

Person Reliability and Item Reliability

To check the stability of persons and items with Rasch reliability values ranging from zero to one which is interpreted as Cronbach's Alpha (William J Boone & Noltemeyer, 2017). Any reliability value close to one can be considered internally consistent (Kam et al., 2011; Maat & Rosli, 2016). Reliability is considered ideal if it is greater than 0.90 (Choi,

Mericle, and Harachi, 2006). In this table, the person reliability index value is 0.85, item reliability is 0.99, and the Cronbach alpha coefficient is 0.87. The high estimates of reliability illustrate that there is a consistent interaction between student responses and items. Thus, the instrument has ideal psychometric internal consistency and is considered a reliable instrument to use.

TABLE 3.1 Data.xls OUTPUT Nov 19 2024 11:29
INPUT: 220 PERSON 36 ITEM REPORTED: 220 PERSON 36 ITEM 2 CATS WINSTEPS 5.3.1.

SUMMARY OF 220 MEASURED PERSON

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	20.5	36.0	.36	.44	.98	-.02	1.06	.15
SEM	.4	.0	.08	.00	.01	.07	.06	.06
P.SD	6.6	.0	1.19	.07	.20	.96	.83	.94
S.SD	6.6	.0	1.19	.07	.20	.96	.84	.94
MAX.	34.0	36.0	4.16	1.00	1.71	3.41	8.74	5.42
MIN.	5.0	36.0	-2.54	.39	.60	-2.23	.25	-1.08

REAL RMSE	.46	TRUE SD	1.10	SEPARATION 2.40	PERSON RELIABILITY .85
MODEL RMSE	.44	TRUE SD	1.11	SEPARATION 2.50	PERSON RELIABILITY .86
S.E. OF PERSON MEAN = .08					

PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .87 SEM = 2.41
STANDARDIZED (50 ITEM) RELIABILITY = .90

SUMMARY OF 36 MEASURED ITEM

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	125.6	220.0	.00	.20	1.00	-.16	1.06	-.14
SEM	8.6	.0	.30	.02	.03	.35	.07	.32
P.SD	51.1	.0	1.75	.14	.17	2.08	.40	1.92
S.SD	51.8	.0	1.78	.14	.17	2.11	.41	1.95
MAX.	211.0	220.0	6.46	1.01	1.38	4.87	2.24	4.27
MIN.	1.0	220.0	-3.38	.15	.76	-3.53	.59	-3.09

REAL RMSE	.25	TRUE SD	1.73	SEPARATION 6.82	ITEM RELIABILITY .98
MODEL RMSE	.25	TRUE SD	1.74	SEPARATION 7.01	ITEM RELIABILITY .98
S.E. OF ITEM MEAN = .30					

ITEM RAW SCORE-TO-MEASURE CORRELATION = -.96
Global statistics: please see Table 44.
UMEAN=.0000 USCALE=1.0000

Person and Item Separation Indeks

Person and Item Separation Index is an estimate of an instrument that can differentiate between student abilities. The greater the person separation index and item separation index means the possibility of the spread of students responding to items correctly and how wide the spread of items is from easy to difficult items (Mez et al., 2012; Perera et al., 2018). The separation index value ranges from zero to infinity, a higher separation value indicates better separation (Linacre, 2012). According to (Duncan et al., 2003) the index criteria with a value of 1.50 is acceptable, 2.00 is good, and 3.00 is very good. The person separation index is 2.40 and the item separation index is 6.82 which provides information about the level of ability in the range of student distribution. Thus, the existence of criteria for students' ability levels supports reliable instruments.

Precision of Measurement

Precision of Measurement is a strong reliability of the instrument and describes the conclusion. Accurate and reliable measurements are very important to evaluate the reliability and strength of an instrument (Perera et al., 2018; Zagorsek & Stough, 2006). A good standard error in an instrument must be less than 0.5 (Pereraet al., 2018). The

estimated item values obtained in the column "Model S.E." equal to 0.25 logits. This can be interpreted as precision of measurement being a reliable indication of item fit. It was concluded that the level of reliability of the instrument was reliable and showed good measurement precession.

CONCLUSION

Overall, the quantitative literacy instrument developed shows an adequate distribution of difficulty levels to measure respondents' abilities in the average to medium ability range. However, there are limitations in covering extreme abilities, both at very low and very high levels, which indicates a gap in measurement coverage. This instrument shows a high degree of reliability and precision, as well as a good ability to separate individuals based on differences in their abilities. In improving the quality of the instrument, it is recommended to add items with a higher level of difficulty and revise items that are too easy or too difficult. This step aims to create an instrument that is more balanced and capable of comprehensively representing respondents' abilities across the entire range of abilities.

REFERENCES

Aryadoust, Vahid. "A review of comprehension subskills: A scientometrics perspective." *System* 88 (2020): 102180.

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and

Boone, W. J., Staver, J. R., Yale, M. S., Boone, W. J., Staver, J. R., & Yale, M. S. (2014). Person reliability, item reliability, and more. *Rasch analysis in the human sciences*, 217-234.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied psychological measurement*, *41*(3), 178-194.

Fisher, D. G., Reynolds, G. L., Jaffe, A., & Johnson, M. E. (2007). Reliability, sensitivity and specificity of self-report of HIV test results. *AIDS care*, *19*(5), 692-696.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and evaluation in Counseling and Development*, *36*(3), 181-191.

Huberty, J., Vener, J., Gao, Y., Matthews, J. L., Ransdell, L., & Elavsky, S. (2013). Developing an instrument to measure physical activity related self-worth in women : Rasch analysis of the Women ' s Physical Activity Self-Worth Inventory. Psychology of Sport & Exercise, 14(1), 111–121. https://doi.org/10.1016/j.psychsport.2012.07.009

Kaliski, P. K., Wind, S. A., Engelhard, G., Morgan, D. L., Plake, B. S., & Reshetar, R. A. (2013). Educational and Psychological Measurement. https://doi.org/10.1177/0013164412468448

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing research*, *35*(6), 382-386.

Mardapi, D., 2008, Teknik Penyusunan Instrumen Tes dan Non Tes. Yogyakarta : Mitra Cendikia Offse

Naga, Dali Santun,. (2013). Teori Sekor pada Pengukuran Mental.Jakarta: PT. Nagarani Citrayasa.

Nitko, J. A. (1996). Educational Assessment of Student. New Jersey: Prentise-Hall.

OECD. (2022). *PISA 2022 Results (Volume I): What Students Know and Can Do*. OECD Publishing. https://doi.org/10.1787/5f7b08b9-en

O'Neill, S. C., & Stephenson, J. (2011). The measurement of classroom management self-efficacy: a review of measurement instrument development and influences. *Educational Psychology*, *31*(3), 261-299.

Poerwanti, E. (2008). Standar penilaian badan standar nasional pendidikan (bsnp).

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in nursing & health*, *29*(5), 489-497. practitioners. Cogent Education, 4(1), 1416898. https://doi.org/10.1080/2331186X.2017.1416898.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.2307/1164671

Reckase, M. D., & Reckase, M. D. (2009). Unidimensional item response theory models. *Multidimensional item response theory*, 11-55.

Rempusheski, V. F., & O'Hara, C. T. (2005). Psychometric properties of the grandparent perceptions of family scale (GPFS). *Nursing Research*, *54*(5), 363-371.

Steen, L. A. (2001). Mathematics and numeracy: Two literacies, one language. *The mathematics educator*, *6*(1), 10-16.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.

Thiagarajan, Sivasailam, dkk. (1974). Instructional Development for Training Teachers of Exceptional Children. Washinton DC: National Center for Improvement Educational System.