VALIDATION AND RELIABILITY OF JUNIOR HIGH SCHOOL STATE EXAMINATION INSTRUMENTS 2015 YOGYAKARTA REGION PACKAGE 1, 2, 3

Aji Joko Budi Pramono[1], Muhlis Malaka[2]

[1]*Universitas Islam Negeri Raden Mas Said*
[2]*Institut Agama Islam Negeri Ternate*

**Corresponding Author: ajijoko@gmail.com*

Received: June 2, 2024
Accepted: July 25, 2024
Available online: Agustus 10, 2024

Abstract *-A good assessment is an assessment that meets the principles of assessment, namely: valid, objective, fair, integrated, open, comprehensive and sustainable, systematic, criteria-based, and accountable. The instrument must meet the requirements of substance, construction, and language, have evidence of validity, and reliability. In this study will be reviewed about the instruments that have been used for the UN in 2015 in Yogyakarta area. The study of instruments in this study includes the suitability of latent constructs, indicators of existing items. The purpose of preparing this instrument is to confirm the latent constructs or basic competencies with the items that have been tested in 2015, this research was conducted using secondary data, namely the 2015 national exam data in the form of responses or answers to student exam results in the Yogyakarta area. The steps taken in this study make indicators, difficulty levels and differentiation of items. content validation or AIKEN and confirm latent constructs and UN 2015 items using Confirmatory Factor Analysis (CFA) techniques. From the Reliability Estimation results, the largest contribution value to the latent variable seen from the CR value is latent variable B, which is equal to 0.9658 contribution, which indicates that the item for a latent variable (Competence) is a reliable indicator in measuring the latent change. The level of difficulty of the items in UN 2015 packages 1, 2 and 3 of Yogyakarta region is best in item no. 2, 4, 6, 8, 10, 11, 16, 21, 38, 39, where the difficulty level value is at a moderate level and the differentiating power of 40 items has a good category all. there are 10 items with sufficient validation and 30 items with high validation, meaning that all items can be used again to make measurements. Based on the results of model fit testing, 5 criteria for model fit show good / fit while 3 criteria are not good / fit all loading factor values have a significant effect (unidimensional) on latent variables in first order Confirmatory Factor Analysis (CFA). The largest contribution to the latent variable seen from the CR value is latent variable B, which is a contribution of 0.9658, which indicates that the item for a latent variable (Competence) is a reliable indicator in measuring the latent change.*
Keyword: *Confirmatory Factor Analysis (CFA), Instrument Validity and Reliability, National Exam Basic Competencies*

1. Introduction

Quality assessment is one that meets the principles of assessment. Permendikbud Number 23 of 2016 states that the principles of assessment are valid, objective, fair, integrated, open, comprehensive and sustainable, systematic, criteria-based, and accountable. What greatly affects the quality of assessment is the assessment instrument used. One of the procedures for assessing the learning process and learning outcomes by teachers is to analyze the quality of the instrument. The instrument must meet substance, construction, and language requirements, have evidence of validity, and reliability. One of the stages of instrument development is analyzing or reviewing the instrument(Mardapi, 2018). The analysis includes qualitative analysis and quantitative analysis.

Qualitative analysis includes substance, construction, language, and validity requirements. For quantitative analysis, the analysis includes level of difficulty, differentiation, functioning of exemptions, and reliability. Quantitative analysis can be done in a conventional way and by using a computer. Conventional analysis is an analysis in which the statistical calculations are done manually. This method has the disadvantage that the process takes longer and is prone to errors. This weakness can be overcome by computer-aided analysis. Analysis in this way is faster because all calculations are carried out by computers and there are very few calculation errors. Dyah(2016) argues that until now many learning outcome instruments have not met the requirements of a good test. One of the things that may be the cause is the teacher's ability to make tests that are still low so that measurements become inaccurate. "A test is an instrument to collect data on participants who respond to questions so that participants can demonstrate the maximum ability and mastery they have (Tenri,2018). Budi (2014) says that "the instrument has a very important function and role in order to determine the effectiveness of the learning process".

The National Exam, commonly abbreviated as UN, is a system of evaluating the standards of primary and secondary education nationally and the quality equality of education levels between regions carried out by the Education Assessment Center, Ministry of Education in Indonesia based on the Law of the Republic of Indonesia number 20 of 2003 which states that in the context of controlling the quality of education nationally, evaluation is carried out as a form of accountability of education providers to interested parties. It further states that evaluations are conducted by independent institutions periodically, thoroughly, transparently and systematically to assess the achievement of national education standards and the evaluation monitoring process must be carried out on an ongoing basis.The evaluation monitoring process is carried out continuously and continuously and will ultimately be able to improve the quality of education. Improving the quality of education begins with determining standards. one of the subjects tested for junior high school (smp) is mathematics where the UN subject instrument has been published through Permendikbud number 37.

This study will examine the instruments that have been used for the UN in 2015 in Yogyakarta. The study of instruments in this study includes the suitability of latent constructs or competency standards, indicators of existing items. The purpose of preparing this instrument is to confirm the latent constructs or basic competencies with the items that have been tested in 2015, so that it can be known the validity, reliability and suitability between the items and the latent constructs and know about the criteria of the items which include the level of

difficulty of the items and the differentiating power. Analysis of the level of difficulty is intended to determine whether the question is classified as easy or difficult. The level of difficulty is a number that shows the difficulty or ease of a question (Arikunto, 1999: 207) difficulty index is classified as Table 1 below,

Table 1. Classification of Level of Difficulty

P-P	Classification
0.00 – 0.29	difficult,
0.30 – 0.69	medium
0.70 – 1.00	easy

(Arikunto; 1999: 210)

Another opinion says that good items in terms of difficulty index are items with moderate difficulty, which are in the range of 0.3 to 0.7 (Mardapi, 2008: 143; Prabowo, 2016: 558). A good item in terms of its differentiating power is an item that has a differentiating power index of more than 0.2 (Fernandes, 1984: 25-29; Prabowo, 2016: 559). The differentiating power of a question is the ability of a question to distinguish between students with high abilities and students with low abilities.

A number of items that have been tested will show how good the differential index between the items is, this differential index will inform the suitability of the test measuring instrument's ability to describe the differences between test takers who have high abilities and those with low abilities. The index of the differential power of this question is obtained from the difference in the proportion of test takers who answer from each group. Thus, the differentiation index can provide an overview of the validity of the question to distinguish between high-ability test takers and low-ability test takers. The negative sign indicates that low ability test takers answered correctly while high ability test takers answered incorrectly. Thus, test questions that have a negative differentiation index indicate the reverse quality of test participants. Items with a negative difference index must be corrected before being used again, or if the difference index is too bad, it should not be used again. The difference index can be calculated by dividing the groups, namely the upper group which is the group of test takers who have high abilities with the lower group, namely the group of test takers who have low abilities The difference index is defined as the difference between the proportion of correct answers in the upper group and the proportion of correct answers in the lower group. (Crocker dan Algina, 1986). Experts divide this group into 27% or 33% of the upper group and 27% or 33% of the lower group. A good item in terms of its differential power is an item that has a differential power index of more than 0.2. Item validity parameters are also very important to ensure that the assessment instrument used actually measures what should be measured. The definition of content validity is the extent to which the elements of the assessment instrument are relevant and represent the construct of the measuring instrument targeted for a particular purpose (Haynes, dkk. 1995) there are three approaches in examining the validity of a measuring instrument, namely 1) content validity, 2) construct validity, and 3) criterion validity (Suryabrata, 2005). To determine this agreement, validity indices can be used, including the index proposed by Aiken (Kumaidi, 2014) . According to Guion (1977) content validity can be determined based on expert justification. The procedures taken to make the test instrument valid are: defining the grid to be measured, determining the grid that will be measured by each question, and comparing each question with the predetermined grid.

Construct validity is a description that shows the extent to which the measuring instrument shows results that are in accordance with the theory (Azwar,

2005). The process of testing construct validity is to link the measuring instrument with other measuring instruments that have similar concepts or with other measuring instruments that are theoretically related (Murphy & Davidshofer, 1991) .This testing process uses Confirmatory Factor Analysis (CFA). CFA is a multivariate analysis method that can be used to confirm whether the measurement model built is in accordance with what is hypothesized. Meanwhile, according to Joreskog and Sorborn (1993) CFA is used to test unidimensionality, validity and reliability of construct measurement models that cannot be measured directly In confirmatory factor analysis, there are latent variables and indicator variables. Latent variables are variables that cannot be formed and built directly while indicator variables are variables that can be observed and measured directly.

2. Methodology

This research was conducted using secondary data, namely the 2015 national exam data in the form of responses or answers to student exam results in the Yogyakarta area. The steps taken in this study are first to make the indicators of the items in accordance with the 2015 UN items, the second step of the item is to find the criteria for the level of difficulty and the differentiation of the items using the response, the third is to validate using the content validation technique or AIKEN and the fourth is to confirm the latent construct and the UN 2015 items using CFA. Data analysis techniques are carried out quantitatively using the R Program to determine the level of difficulty of the questions and the differentiation of the questions, Lisrel 3.0 for significance testing using CFA and AIKEN.

3. Results

Based on Permendikbud number 37 of 2018, there are 5 basic competencies and 24 latent competency indicators. There are 3 packages of question items for the Yogyakarta region, namely package 1, package 2 and package 3. Each package contains 40 items, each item uses multiple choice answers 4 alternative answers The questions that have been written by the participants are then reviewed. The review carried out is a qualitative review which includes the suitability of the material, the construction of the questions, and the language used. After the review has been carried out, then each item in the question package is made an indicator commonly referred to as the item indicator, the next step is that all item indicators are adjusted to the basic / latent competency indicators so that they are presented in the instrument table as in table 1.

Table 1. UN 2015 instrument

No	COMPETENCY	INDICATOR	QUESTION ITEM INDICATOR	Package 1 Question	Package 2 Question	Package 3 Question
1	Using concepts of arithmetic operations and properties of numbers, ratio of numbers, powers, roots, social arithmetic, number sequences, and their use in problem-solving.	Solving problems related to addition, subtraction, multiplication, or division operations on numbers.	Given a number of questions where each correct, incorrect, and unanswered answer is scored, a student answers a number of questions correctly and incorrectly, calculate the score obtained.	1	1	1
		Solving problems related to ratios	Given two problems with equivalent component ratios, students can calculate using the ratio pattern.	2, 3	2, 3	
		Solving problems related to operations with exponents or roots	Given problems related to operations with exponents	4, 5	4, 5	

			or roots, students solve the operations.			
		Solving problems related to banking or cooperatives in simple social arithmetic.	Given a banking problem involving simple arithmetic, students calculate one of the banking issues.	6	6	
		Solving problems related to number sequences and series	Students calculate the nth term of a number sequence.	7, 8, 9	7, 8, 9	
2	Understanding algebraic operations, concepts of linear equations and inequalities, line equations, set relations, functions, linear equation systems, and their use in problem-solving.	Determining the factoring of algebraic forms.	Students calculate the result of factoring algebraic forms.	10	10	10
		Solving problems related to linear equations or linear inequalities of one variable	Given forms of linear equations or inequalities, students solve the equations.	12, 13	12, 13	
		Solving problems related to sets	Given a specific set, students solve the set of solutions for a linear equation.	11, 14, 15	11, 14, 15	
		Solving problems related to functions	Given a specific function equation, students calculate the value of the function whose equation is known.	16	16	
		Determining the gradient of a line equation or its graph	Given an equation, students calculate the gradient, create the equation, or graph.	17, 18	17, 18	
		Solving problems related to linear equations of two variables	Given the coefficient values from equations involving x and y, students calculate the coefficient values of x or y.	19, 20	19, 20	
3	Understanding the concept of similarity, properties, and elements of plane shapes, and the concept of angle or line relationships, and using them in problem-solving.	Solving problems using the Pythagorean theorem	Given the hypotenuse and one side of a triangle, students calculate the length of the other side.	21	21	21
		Solving problems related to the area of plane shapes	Given the side lengths of a plane shape, students calculate the area of the plane shape.	22, 23	22, 23	
		Solving problems related to the perimeter of plane shapes	Given the side lengths of a plane shape and other components, students calculate the perimeter.	24	24	
		Solving problems related to similarity or congruence	Given a plane shape with certain similarities, students calculate the similarity of the other sides.	25, 26	25, 26	
		Solving problems related to the relationship between two lines: angle size (complementary or supplementary)	Given two lines with different lengths in a case, students calculate the length of one line if the length of the other line changes.	27	27	
		Solving problems related to special lines in triangles	Given the size of an angle with a specific complement, students calculate the supplementary angle or distinguish congruent lines.	28, 29	28, 29	
		Solving problems related to elements/parts of circles or the relationship between two circles	Given two circles with the length of the chord and the radius of one circle, students calculate the radius of the other circle.	30	30	
		Understanding the properties and elements of solid figures and using them in problem-solving.	Determining the elements of solid figures	31	31	31
		Solving problems related to the	Given the framework of a solid figure with the length of one edge known,	32	32	

		framework or nets of solid figures	students calculate the maximum number of other frameworks.			
		Solving problems related to the volume of solid figures	Given a solid figure such as a half-sphere with a specified diameter, students calculate the volume of the object.	33, 34	33, 34	
		Solving problems related to the surface area of solid figures	Given a solid figure with sides, students calculate the surface area.	35, 36	35, 36	
4	Understanding statistical concepts and applying them in problem-solving.	Determining measures of central tendency or using them in solving everyday problems.		37	37	37
		Solving problems related to data presentation or interpretation		38, 39	38, 39	
5	Understanding the concept of probability of an event and applying it in problem-solving.	Solving problems related to the probability of an event		40	40	40

Table 2. Content validity results

Item	**Penilai**					**s1**	**s2**	**s3**	**s4**	**s5**	**∑s**	**V**	**Ket**
	I	**II**	**III**	**IV**	**V**								
Item_01	3	3	3	3	3	2	2	2	1	2	9	0,600	Medium
Item_02	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_03	4	3	4	4	3	3	2	3	3	2	13	0,867	High
Item_04	3	3	3	3	3	2	2	2	2	2	10	0,667	Medium
Item_05	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_06	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_07	4	3	3	3	3	2	2	2	2	2	10	0,667	Medium
Item_08	3	4	3	4	3	2	3	2	3	2	12	0,800	Medium
Item_09	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_10	3	4	4	4	3	2	3	3	3	2	13	0,867	High
Item_11	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_12	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_13	3	4	4	3	3	2	3	3	2	2	12	0,800	Medium
Item_14	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_15	4	4	3	3	3	3	3	2	2	2	12	0,800	Medium
Item_16	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_17	4	3	4	4	3	3	2	3	3	2	13	0,867	High
Item_18	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_19	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_20	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_21	3	4	4	3	3	2	3	3	2	2	12	0,800	Medium
Item_22	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_23	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_24	4	3	3	3	3	3	2	2	2	2	11	0,733	Medium
Item_25	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_26	3	4	3	4	3	2	3	2	3	2	12	0,800	Medium
Item_27	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_28	4	3	4	4	3	3	2	3	3	2	13	0,867	High
Item_29	4	4	3	4	3	3	3	2	3	2	13	0,867	High
Item_30	3	4	4	3	4	2	3	3	2	3	13	0,867	High
Item_31	4	4	3	4	4	3	3	2	3	3	14	0,933	High
Item_32	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_33	4	3	4	4	3	3	2	3	3	2	13	0,867	High
Item_34	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_35	4	4	4	4	4	3	3	3	1	3	13	0,867	High
Item_36	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_37	3	3	3	4	3	2	2	2	3	2	11	0,733	Medium
Item_38	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item_39	4	4	4	3	3	3	3	3	2	2	13	0,867	High
Item_40	4	4	4	4	3	3	3	3	3	2	14	0,933	High
Item 1-40	**151**	**151**	**142**	**146**	**123**	**110**	**111**	**102**	**103**	**83**	**509**	**0,848**	

The UN results data that have been carried out in the Yogyakarta area are then analyzed. The analysis was carried out with the ITEMAN analysis technique using MS Exel to determine the characteristics of the test items that had been used. The results of the analysis displayed include the level of difficulty and differentiation of the questions, then categorize the level of difficulty and differentiation of the question

Content Validity

The content validity test in this study used AIKEN validity. The experts involved in the process of assessing the content validity of the UN instrument consisted of two mathematics lecturers and 3 mathematics teachers. The experts gave an assessment of the suitability between the items and the indicators using a Likerd scale (Score 1: Invalid, Score 2: Less Valid, Score 3: Quite Valid, Score 4: Valid, Score 5: Very Valid). Furthermore, the results are interpreted, if the agreement index is less than 0.4 then it is said to be low validation, if between 0.4 - 0.8 it is said to be moderate validation and if more than 0.8 is said to be high Heri Retnawati, (2015) The results of content validity using AIKEN can be seen in table 2 as follows:

Based on table 1, it can be seen that there are 10 items with sufficient validation and 30 items with high validation.

Level of Difficulty

The test instrument items can be said to be good if the test items have a level of difficulty in the interval 0.31-0.70. This shows that the item is not too difficult and also not too easy. Analysis of the level of difficulty of each item is shown in the following table 5.
in table 5 below

Table 5. Level of difficulty of question items

Question	Difficulty Level	Category	Question	Difficulty Level	Category
Item 1	-0.90	Low	Item 21	0.45	Medium
Item 2	0.62	Medium	Item 22	0.75	High
Item 3	-1.13	Low	Item 23	1.79	High
Item 4	0.62	Medium	Item 24	0.07	Low
Item 5	-0.10	Low	Item 25	1.25	High
Item 6	0.30	Medium	Item 26	-1.25	Low
Item 7	-1.07	Low	Item 27	-0.33	Low
Item 8	0.69	Medium	Item 28	1.31	High
Item 9	1.37	High	Item 29	0.73	High
Item 10	-0.30	Low	Item 30	1.14	High
Item 11	0.67	Medium	Item 31	0.79	High
Item 12	1.10	High	Item 32	0.25	Low
Item 13	-0.81	Low	Item 33	0.20	Low
Item 14	-0.53	Low	Item 34	-0.03	Low
Item 15	0.25	Low	Item 35	0.73	High
Item 16	0.30	Medium	Item 36	0.82	High
Item 17	0.88	High	Item 37	-1.74	Low
Item 18	0.02	Low	Item 38	0.67	Medium
Item 19	-0.84	Low	Item 39	0.49	Medium
Item 20	1.42	High	Item 40	0.88	High

Based on the results of the analysis of question items in terms of the level of difficulty of the UN mathematics question items in Yogyakarta, it is known that of the 40 questions that have been tested, there are 17 questions in the low difficulty category, 10 questions in the Medium category, and 13 questions in the high difficulty category.

item differentiation

Analysis of the item differential of the question items using the R program, the

differential itemcof the question items is then categorized as good if the question items have a differential index of more than 0.2 and the category is not good if the question items have a item differential index of less than 0.2. The results of the R program analysis are presented in table 3 below:

Table 3. item differentiation

Item	Discrimination Index	Category	Item	Discrimination Index	Category
Item 1	1.28	Good	Item 21	1.28	Good
Item 2	0.97	Good	Item 22	0.55	Good
Item 3	0.95	Good	Item 23	0.51	Good
Item 4	1.25	Good	Item 24	2.13	Good
Item 5	1.51	Good	Item 25	0.35	Good
Item 6	1.02	Good	Item 26	1.08	Good
Item 7	0.86	Good	Item 27	0.92	Good
Item 8	0.56	Good	Item 28	0.39	Good
Item 9	0.29	Good	Item 29	0.67	Good
Item 10	0.89	Good	Item 30	0.72	Good
Item 11	0.39	Good	Item 31	0.57	Good
Item 12	0.59	Good	Item 32	0.44	Good
Item 13	0.86	Good	Item 33	0.64	Good
Item 14	0.76	Good	Item 34	0.76	Good
Item 15	0.87	Good	Item 35	0.40	Good
Item 16	0.64	Good	Item 36	0.22	Good
Item 17	0.89	Good	Item 37	0.72	Good
Item 18	0.77	Good	Item 38	0.78	Good
Item 19	1.03	Good	Item 39	1.22	Good
Item 20	0.30	Good	Item 40	0.71	Good

Construct Validity

Construct validity is an instrument that shows the extent to which the instrument reveals a theoretical trait or construct that it wants to measure. In this case the construct is the framework of a concept. This notion of construct is latent and abstract so that it is related to many empirical indicators that require analytical tests such as confirmatory factor analysis. From Table 1. The instrument table is then tested using the CFA technique with the help of lisrel software. To facilitate testing, it will be tested between latent variables (basic competencies) and question items that have been adjusted to their indicators with each latent variable (competency standard) symbolized by letters A, B, C, D and E. Medium question items are symbolized by following the letters of each latent variable. The image of the test results with Lisrel is in Figure 1 (standardized image) and Figure 2 (T value). Testing is done using student response data that has been tested.

The Fit Model results are in table 4.

Table 4. Model fit table of the UN instrument

Indikator	Nilai Patokan	Nilai Perolehan	Kriteria Model Fit
Chi-square	< 2df	700	Tidak fit
Signifikansi (p-value)	≥ 0.05	0,00001	Tidak fit
RSMEA	≤ 0.08	0.026	fit
Goodness of Fit Index (GFI)	≥ 0.90	0.91	fit
Adjusted Goodness of Fit Index (AGFI)	≥ 0.90	0.92	fit
Normed Fit Index (NFI)	≥ 0.90	0.91	fit
Comparative Fit Index (CFI)	≥ 0.95	0.87	Tidak fit
Incremental Fit Index (IFI)	≥ 0.95	0,92	fit

To determine the validity of the UN instrument, it can be seen in the loading factor value as shown in table 5 as follows.

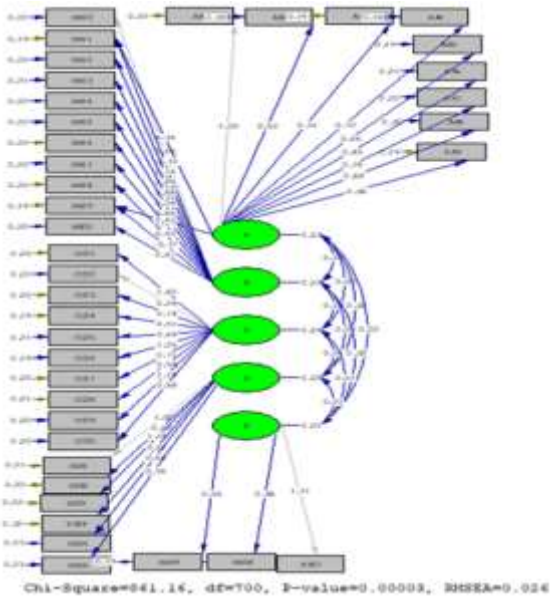

Figure 1. Standardized picture of UN instrument

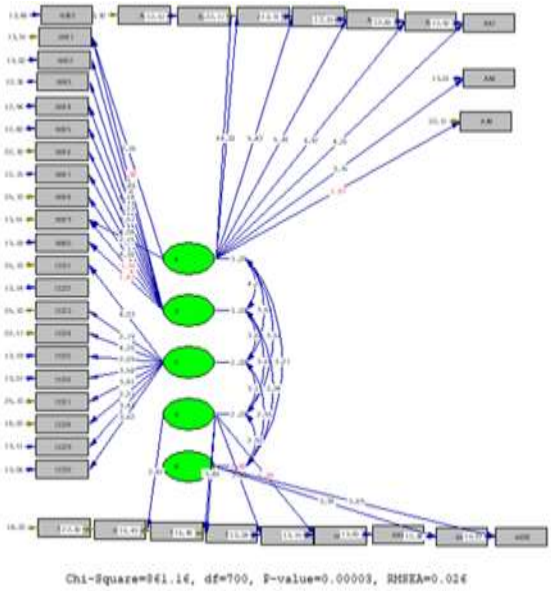

Figure 2. T value of UN instrument

Table 5. Loading factor table

Competency	Item Question	Loading > 0,5 (Standarized)	Decision	Loading > 1,96 T-Value	Decesion
A	Item1	0,37	Invalid	1	Invalid
	Item2	0,52	Valid	4,93	Valid
	Item3	0,57	Valid	4,33	Valid
	Item4	0,72	Valid	5,42	Valid
	Item5	0,56	Valid	5,41	Valid
	Item6	0,55	Valid	4,97	Valid
	Item7	0,78	Valid	4,21	Valid
	Item8	0,68	Valid	3,75	Valid
	Item9	0,66	Valid	1,97	Valid
B	Item10	0,14	Invalid	1	Invalid
	Item11	-0,13	Invalid	-1,92	Invalid
	Item12	0,76	Valid	3,86	Valid
	Item13	0,89	Valid	4,18	Valid
	Item14	0,88	Valid	4,13	Valid
	Item15	0,24	Invalid	4,57	Valid
	Item16	0,86	Valid	4,08	Valid
	Item17	1,07	Valid	2,15	Valid
	Item18	0,93	Valid	4,36	Valid
	Item19	-0,77	Invalid	-1,51	Invalid
	Item20	0,67	Valid	1,67	Valid
C	Item21	1,82	Valid	1	Invalid
	Item22	0,24	Invalid	4,03	Valid
	Item23	0,74	Valid	2,79	Valid
	Item24	0,51	Valid	4,20	Valid
	Item25	0,64	Valid	2,09	Valid
	Item26	0,71	Valid	3,58	Valid
	Item27	0,56	Valid	3,81	Valid
	Item28	0,54	Valid	2,27	Valid
	Item29	0,68	Valid	3,43	Valid
	Item30	0,51	Valid	3,62	Valid
D	Item31	0,19	Invalid	1	Invalid
	Item32	0,67	Valid	2,87	Valid
	Item33	0,63	Valid	3,63	Valid
	Item34	0,67	Valid	3,83	Valid
	Item35	0,17	Invalid	2,40	Valid
	Item36	0,86	Valid	1,59	Valid
E	Item37	0,22	Tidak Valid	1	Invalid
	Item38	0,53	Valid	3,39	Valid
	Item39	0,56	Valid	3.69	Valid

it is necessary to test the reliability construct, invalid question items are no longer used, the factor loading coefficient value. This construct reliability can be estimated after the researcher proves the validity of the construct with confirmatory factor analysis until obtaining a suitable model (fit model) Heri retnawati (2016). Reliability Estimation CR uses the factor loading of each indicator that composes the instrument (λ) and the unique error index of each indicator.

(ξ). With the following formula (Geldhof, Preacher, Zyphur, 2014).

$$CR = \frac{\left(\sum_{i=1}^{i} \lambda_i\right)^2}{\left(\sum_{i=1}^{i} \lambda_i\right)^2 + \left(\sum_{i=1}^{i} \delta\right)}$$

The greater this CR value, it shows that the divide of a latent variable (Competence) is a reliable indicator in measuring these latent changes. According to Hair et al. (2010), the CR value that is still acceptable is ideally more than 0.7. Based on Figure 1, the CR value can be calculated as follows.

CR Competency A

Competency	Loading Factor	Error
B	0,76	0,22
	0,89	0,21
	0,88	0.19
	0,86	0,22
	1,07	0,25
	0,93	0,19
	0,67	0,21
Total	6,06	1,30
Total ²	36,7236	
CR	0,965810707	

CR Competency B

Competency	Loading Factor	Error
B	0,76	0,22
	0,89	0,21
	0,88	0.19
	0,86	0,22
	1,07	0,25
	0,93	0,19
	0,67	0,21
Total	6,06	1,30
Total ²	36,7236	
CR	0,965810707	

CR Competency C

Competency	Loading Factor	Error
D	0,53	0,21
	0,56	0,19
Total	1,09	0,40
Total ²	1,1881	
CR	0,748126692	

CR Competency D

Competency	Loading Factor	Error
D	0,67	0,24
	0,63	0,19
	0,67	0,18
	0,17	0,22
	0,86	0,2
Total	3	1,03
Total ²	9	
CR	0,897308076	

CR Competency E

Competency	Loading Factor	Error
D	0,53	0,21
	0,56	0,19
Total	1,09	0,40
Total ²	1,1881	
CR	0,748126692	

Summary of all CR tables is shown in Table 5 below.

Table 5. Summary of CR values

No	Competency	CR	decision
1	A	0,940395978	Reliabel
2	B	0,965810707	Reliabel
3	C	0,959713604	Reliabel
4	D	0,897308076	Reliabel
5	E	0,748126692	Reliabel

CR > 0,7 , Reliabel

From the results of the CR analysis presented in table 5. that all loading factors have a CR value of more than 0.7 so that they can be said to be reliable.

4. Discussion

The validation and reliability analysis of the Junior High School National Examination (UN) instruments in Yogyakarta for 2015 provides significant insights into the quality and effectiveness of these assessment tools. This study examines various aspects, including content validity, difficulty level, item discrimination, and construct validity, analyzed through qualitative and quantitative methods. Content validity ensures that the assessment instruments adequately cover the curriculum and learning objectives they are intended to measure (Haynes, Richard, & Kubany, 1995). In this study, content validity was assessed using Aiken's V method, involving experts from the field of mathematics education. The results indicated that out of 40 items, 30 items were categorized as having high validity, while the remaining 10 had moderate validity. This suggests that most items are well-aligned with the intended learning outcomes (Retnawati, 2016). The high content validity of these items underscores their appropriateness in assessing students' competencies in mathematics, aligning well with the educational goals outlined by curriculum standards (Suryabrata, 2005). Furthermore, this alignment highlights the effectiveness of these items in providing a comprehensive evaluation of students' understanding and mastery of mathematical concepts. Difficulty level of test items is a critical factor in determining their effectiveness. Items that are too difficult or too easy can skew the results and fail to accurately measure student competencies (Arikunto, 1999). The analysis of the difficulty level in the UN test instruments showed that out of the 40 items tested, 42.5% of the items were categorized as low difficulty, 25% as medium, and 32.5% as high difficulty. This distribution indicates a range of difficulty levels, with the highest proportion of items in the low difficulty category, followed by high and then medium categories. Thus, it can be said that the overall difficulty level of the items is not well-balanced. Good items are those that are neither too easy nor too difficult. If the items are too easy, both high-achieving and low-achieving students can solve them. Conversely, if the items are too difficult, neither high-achieving nor low-achieving students can solve them. As a result, items that are too easy or too difficult cannot differentiate between high-achieving and low-achieving students. Therefore, an item is considered good if its difficulty level ranges from 0.31 to 0.70, categorized as medium (Mardapi, 2008; Prabowo, 2016).

Item discrimination relates to an item's ability to differentiate between students who understand the material and those who do not (Crocker & Algina, 1986). Positive discrimination values indicate high discrimination power, while negative values indicate low discrimination power. The analysis revealed that overall, the items had discrimination values greater than 0.2, meaning the items were accepted or good at differentiating between students who understand the material and those who do not. Items with good discrimination should be included in the item bank due to their quality and can be used again in future tests (Fernandes, 1984; Prabowo, 2016). Construct validity was assessed using Confirmatory Factor Analysis (CFA), which confirmed the alignment of the items with the underlying constructs they were intended to measure. The fit indices from the CFA indicated a good model fit for most

criteria, although some indices did not meet the desired thresholds. For instance, the Chi-square and significance p-value did not fit well, but the RMSEA, GFI, AGFI, NFI, and IFI showed good fit. Goodness-of-Fit indices showed that the Goodness-of-Fit Index (GFI), Adjusted Goodness-of-Fit Index (AGFI), and Incremental Fit Index (IFI) had values closer to 1.000 (GFI=0.91, AGFI=0.92, IFI=0.92). Comparative Fit Index (CFI), Relative Fit Index (RFI), and Tucker-Lewis Index (TLI) had values approaching 0.87 (CFI=0.87, IFI=0.92, TLI=0.965). Root Mean Square Error of Approximation (RMSEA) < 0.08 (=0.026). Based on the model fit test results in Table 4, five criteria showed good/fit, while three criteria were not good/fit. These results indicate that although some criteria did not fit, the values closer to one indicate a good model and validated. Based on Table 3, it can be seen that the standardized loading factor value must be greater than 0.5. According to Hair et al. (2010), an acceptable factor loading value is more than 0.5, and when it is equal to 0.7 or above, it is considered good for one indicator. For Competency A, out of 9 items, 8 were valid, and 1 was invalid. Competency B consisted of 11 items with 7 valid items and 4 invalid items. Competency C consisted of 10 items with 9 valid items and 1 invalid. Competency D consisted of 6 items with 4 valid items and 2 invalid items. Competency E consisted of 3 items with 2 valid items and 1 invalid item. Invalid items should be dropped and not used again in future assessments. Based on the latent construct model fit test results, the indicators of the basic competency latent variables showed that all loading factor values significantly influenced (unidimensional) the latent variables in the first-order Confirmatory Factor Analysis (CFA). The largest contribution to the latent variable, seen from the CR value, is latent variable B, with a contribution of 0.9658, indicating that items for a latent variable (competency) are reliable indicators in measuring these latent changes.

5. Conclusion

From the above discussion, we can derive the important understanding that the preparation of test items requires special attention to several aspects to achieve the objectives of the measurement. One crucial aspect is the item discrimination and difficulty level. Based on the analysis of item difficulty levels for the 2015 UN packages 1, 2, and 3 in the Yogyakarta region, it was found that the best items were numbers 2, 4, 6, 8, 10, 11, 16, 21, 38, and 39, where their difficulty levels were at the medium level. Additionally, the discrimination power of the 40 items showed good categories, indicating that these items could distinguish between high and low ability students. Referring to the ITEMAN analysis, there were 10 items with sufficient validation and 30 items with high validation, which means that all items can be used again for measurement. These results affirm that most items have high content validity, appropriate difficulty levels, and good discrimination power. Based on the model fit testing results, five model fit criteria showed good or fit, while three criteria did not fit. Although some criteria did not fit, the values close to one indicate that the model is overall good and validated. The latent construct model fit test showed that all loading factor values significantly (unidimensionally) influenced the latent variables in the first-order Confirmatory Factor Analysis (CFA). The largest contribution to the latent variable, seen from the Composite Reliability (CR) value, was latent variable B with a contribution of 0.9658, indicating that the item for a latent variable (competence) is a reliable indicator in measuring latent changes. Thus, the validity and reliability analysis of the 2015 Junior High School National Examination instruments in Yogyakarta confirms the effectiveness of these assessment tools in measuring

students' competencies in mathematics. However, continuous refinement of these instruments is necessary to maintain their validity and reliability, ensuring they remain robust tools for evaluating student performance and informing educational practices.

References

Dyah, F. W., & Putra, A. P. (2016). Pengembangan Instrumen Tes Standar Kognitif pada Mata Pelajaran IPA Kelas 7 SMP Di Kabupaten Banjar. In Proceeding Biology Education Conference: Biology, Science, Enviromental, and Learning, 13(1).

Mardapi, D. (2018). Teknik Penyusunan Instrumen Tes Dan Non Tes. Pendidikan Matematika, 1311440001.

Arikunto, S. (1999). Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: Rineka Cipta.

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart, and Winston.

Tenri Batari, Nursalam Nursalam, Andi Dian Angriani.2018. ”Pengembangan Instrumen Tes Untuk Mengukur Kemampuan Koneksi Matematis”.Jurnal Pendidikan Dasar Islam Vol 5, No 1 (2018)

Budi Manfaat, Siti Nurhairiyah.2014.” Pengembangan Instrumen Tes Untuk Mengukur Kemampuan Penalaran Statistik Mahasiswa Tadris Matematika”.Jurnal Teknologi Informasi, Volume 5 Nomor 2, Oktober. 2014

Undang-Undang Republik Indonesia nomor 20 tahun 2003

Permendikbud nomor 37 tahun 2018.tentang perubahan atas peraturan menteri pendidikan kebudayaan nomor 24 tahun 2016 tentang kompetensi inti dan kompetensi dasar pelajaran pada kurikulum 2013

Rawuh, D. (2014). Kajian Kemampuan Guru Biologi Sman Kabupaten Pringsewu Dalam Menyusun Perangkat Instrumen Penilaian Pada Tahun Ajaran 2011/2012.

Prabowo, Anggit. (2016). Analisis Butir Soal Ujian Akhir Semester Mata Kuliah Analisis Kurikulum dan Materi Pembelajaran Matematika SMA. Prosiding Seminar Nasional Pendidikan Berkemajuan dan Menggembirakan. Medan, 3 Agustus 2016.

Fernandes, H. J. X. (1984). Testing and Measurement. Jakarta: Planning, Evaluation, and Development.

Crocker, L. & Algina, J. 1986. Introduction to Classical and Modern Test Theory.New York: Holt, Rinehart and Winston, Inc.

Haynes, S. N., Richard, D. C., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. Psychological Assessment, 7, 238 - 247.

Suryabrata, S. (2005). Pengembangan Alat Ukur Psikologis. Yogyakarta: Penerbit Andi.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. Educational and psychological measurement, 40(4), 955-959.

Kumaidi. (2014). Validitas dan pemvalidasian instrumen penilaian karakter. Makalah disampaikan dalam Seminar Nasional Pengembangan Instrumen Penilaian Pendidikan Karakter yang valid, diselenggarakan Fakultas Psikologi, Universitas Muhammadiyah Surakarta, 24 Mei 2014

Guion, R.M. 1977. Content Validity, The Source of My Discontent, Applied Psychological Measurement, 1.1-10

Azwar, S. (2005). Dasar-Dasar Psikometri. Yogyakarta: Pustaka Pelajar.

Murphy, K. R., & Davidshofer, C. O. (1991). Psychological Testing: Principles and Applications. New Jersey: Prentice Hall.

Joreskog, K., & Sorborn, D. (1993). LISREL 8: manual. Scientific Software, Mooresville

Retnawati, H. (2016). Validitas reliabilitas dan karakteristik butir. Yogyakarta: Parama Publishing.