Pegem Journal of Education and Instruction, Vol. 13, No. 3 , 2023 (pp. 386-394)

RESEARCH ARTICLE WWW.PEGEGOG.NET

Exploration of Student' Cognitive Mathematics Ability Diagnostic Instruments: Validity, Reliability, and Item Characteristics

Wahyu Hartono[1], Samsul Hadi [2]*, Raden Rosnawati [3], Heri Retnawati [4]

[1]Doctorate Program in Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia

[1]Mathematics Education Department, Universitas Swadaya Gunung Jati, Indonesia

[2-4]Universitas Negeri Yogyakarta, Indonesia

Abstract

Researchers design diagnostic assessments to measure students' knowledge structures and processing skills to provide information about their cognitive attribute. The purpose of this study is to determine the instrument's validity and score reliability, as well as to investigate the use of classical test theory to identify item characteristics. The data used in the form of responses to elementary school mathematics subject. There are 166 students from 5 public elementary schools who participate in the study. The data analysis technique used is the analysis of item characteristics based on classical test theory using the R software package. The results showed that the developed mathematical ability diagnostic instrument had high content validity based on the Aiken formula and valid construct validity based on the CFA approach. According to the Spearman-Brown formulation, the correlation coefficient is about 0.889, indicating high internal consistency reliability. In the index of difficulty level, overall, it is categorized as moderate items. The discriminatory index shows that there are two items, namely items 9 and 17, with low discriminating power, so the two items are not used. Of the 60 total distractors, 5 (8.3%) did not function well because less than 5% of the participants chose them. In contrast, as many as 55 distractors (91.7%) have functioned well.

Keywords: Cognitive Diagnostic, Distractor, Items Characteristic.

Introduction

Mathematics is an ordered subject, which means that students can master a higher level if they have previously mastered a lower level. Because mathematics is intrinsically cumulative, mastery of prerequisite skills is critical (Levine et al., 1992). Cumulative implies that knowledge and skills must accumulate over time. Students whose initial mastery of mathematics subjects is incomplete will find it difficult to follow the learning material.

A cognitive diagnostic assessment is a tool to determine the incompleteness the prerequisite knowledge. Cognitive diagnostic assessments identifying cognitive strengths and weaknesses by assessing students' specialized knowledge structures and processing skills (Leighton & Gierl, 2007). The results of diagnostic assessments are extremely useful in the early management of students in order to minimize learning barriers or optimize learning outcomes. The early treatment will be determined by the diagnostic results. Students with limited abilities, for example, require special consideration; otherwise, they will struggle at the next level of mathematics.

Another goal of the diagnostic test for students with mathematics learning difficulties is to provide teachers with information about students' mental processes and skills. This information will assist teachers in developing effective learning strategies that directly address student issues.

Learning theory is closely related to the applied learning strategy. The four stages of learning theory are comprehension, acquisition, storage, and retrieval (Tall & Razali, 1993). Students with high abilities can complete all four phases, but students with low abilities will have difficulty completing the fourth phase (retrieval). These four stages are critical in the development of students' knowledge and thought processes. Knowledge is structured or built up in a student's mind when he attempts to organize his new experience based on the existing cognitive framework (Bodner, 1986).

Diagnostic learning difficulties, particularly in mathematics, can detect phase imperfections in knowledge construction. Students will benefit from diagnostic testing to identify learning difficulties in mathematics. Conclusions drawn from diagnostic test results are very logical and reasonable. By getting to the root of the problem, diagnostic test results can significantly increase the likelihood of providing students with the most effective and timely instruction. Teachers can use diagnostic test results to correct students' misconceptions and replace incorrect strategies (Leighton & Gierl, 2007).

Corresponding Author e-mail: samsul.hd@gmail.com

https://orcid.org: 0000-0003-3437-2542

How to cite this article: Hartono W, Hadi S*, Raden Rosnawati, Retnawati H. Exploration of Student' Cognitive Mathematics Ability Diagnostic Instruments: Validity, Reliability, and Item Characteristics Pegem Journal of Education and Instruction, Vol. 13, No. 3, 2023, 386-394

Source of support: None.

Conflict of Interest: Nil..

DOI: 10.47750/pegegog.13.03.39

Received: 09.08.2022

Accepted: 27.01.2023 **Publication:** 01.07.2023

Diagnostic test administration must ensure that test results accurately describe students' abilities. The assessment is considered accurate if the results contain a minor possible error. The test instruments, in this case, the basic mathematics (Fractions, Comparison of two quantities, and Scale) diagnostic test questions, must be valid, reliable, and have appropriate item parameters in order to produce accurate results that describe students' abilities. Two approaches can be used to estimate item parameters for this purpose: item response theory and classical test theory. Classical test theory is well known to have flaws. The most notable weakness of classical test theory is the inability to separate examinee and test characteristics (Hambleton, Swaminathan, & Rogers, 1991). It implies that the test only determines the performance of the test taker (student). High-level difficulty test questions make students appear to have low abilities. Simultaneously, questions with a low level of difficulty make students appear highly capable. In other words, the subject/test taker has a strong influence on item parameters and vice versa. As a result, classical test theory cannot be used as a standard because the assessment results are highly dependent on the test taker's subjects. Indeed, many researchers still use classical test theory in their daily work for a variety of reasons. This study will look into the instrument's validity as well as the use of classical test theory to identify the characteristics of items and their benefits.

The early twentieth century saw the emergence of classical test theory. It grew out of the mingling of three remarkable achievements from the previous 150 years: the recognition of measurement errors, the concept of error as a random variable, and the concept of correlation and how to index it (Traub, 2005). According to Sheng (2019), a statistical model at the heart of classical theory relates the observed score to a person's actual score, which is hypothetically the average of all the observed scores that this person would get on the same test after taking it an unlimited number of times. To that end, the CTT package R (Willse, 2018) was recently developed, providing routines for assessing test items and performing general classical theory analyses. Furthermore, the estimation of the CTT measurement error is sample dependent because the standard error of measurement (SEM) can only be estimated using information from a group of respondents (Sharkness & DeAngelo, 2011).

Classical test theory frequently relies on two main statistics to evaluate a single item: item difficulty and item discrimination. The difficulty of the item determines its mean (location). Item discrimination, on the other hand, refers to how much items differ between people based on measured characteristics. For dichotomously scored items, the proportion of people responding in the specified direction (usually denoted by the item's p-value) can be used to calculate the difficulty level. In contrast, a discrimination index can assess item discrimination (Sheng, 2019).

Aside from estimating the level of difficulty and item discrimination, another important aspect of item analysis for multiple-choice items is distractor analysis, which examines how people choose the item for each item. Participants in distractor analysis are classified into trait grades based on their total test scores. In general, we expect a smaller proportion of people with higher trait levels to switch and a greater proportion of people with lower trait levels to switch. A good distraction should ideally attract people in equal parts. If a distraction is unattractive or attracts a tiny proportion of the other distractions, or if the distraction is more attractive than the correct choice for those with a higher trait level, it should be removed or revised (Sheng, 2019). According to Sharkness & DeAngelo (2011), CTT scale scores and their interpretations are always contextual; in particular, they are item and sample-specific. Examinees will have lower correct scores on complex tests and higher correct scores on more accessible tests. However, their ability scores remain constant against any test that might be made to measure the construct (Hambleton, 2005). After investigating the characteristics of the items, now researchers turn to validity and reliability.

In classical test theory, predictive validity is defined as the usual Pearson correlation between the test score and the validation criteria score. This definition serves some purposes. Others, on the other hand, are hampered because the derivation of well-known equations relating to validity and reliability requires independent assumptions of uncorrelated measurement error (Zimmerman, 1998).

Content validity, concurrent validity, predictive validity, and construct validity are the four traditional definitions of validity (Andrich & Marais, 2019). According to Retnawati (2016), validity is related to measurement accuracy and will demonstrate the support of empirical facts and theoretical reasons for the interpretation of test scores or instrument scores. Content validity is determined by experts by determining whether the content is relevant, i.e., taking into account the operational definition of nature. Concurrent validity is established when the results of one instrument are related to the expected results of another relevant instrument. The predictive validity of an instrument is established by relating its results to the exact nature's future performance. Construct validity is demonstrated when the instrument's results match the expectations of a theoretical understanding of the trait. According to Messick (1989), construct validity is an overarching concept, and the other three so-called forms of validity are types of evidence for construct validity. As a result, validity is thought to be synonymous with construct validity.

This study uses content and constructs validity to determine whether or not the developed elementary school mathematical ability diagnostic instrument is valid. Content validity depends on the extent to which empirical

measurements reflect a particular content domain. For example, a test in arithmetic operations will not be valid if the test questions only focus on addition. It ignores subtraction, multiplication, and division (Kurian, 2014). The meaning of content validity is the extent to which the elements in a measuring instrument are genuinely relevant and represent a representation of the contract under the measurement objectives (Haynes, 1995). However, content validity alone is not enough. Logical and empirical evidence about the quality of judgments is a necessary component of validation (Lovitt, 1993). Empirical evidence will be analyzed using construct validity. According to Messick (Lovitt, 1993), proper validation - namely, construct validation - is concerned with tests and the justification of interpretation of test responses and scores. Construct validity was analyzed using Confirmatory Factor Analysis. After investigating the validity, the reliability score will then be investigated.

Score reliability will be calculated using Cronbach value and the Spearman-Brown formulation. The Spearman-Brown formula can be used to determine the dependability of dichotomous or polytomous items. The formula divides the test into odd and even categories (Azwar, 2012). The results of the Spearman-Brown formula were compared with Guilford's reliability coefficient categorization. The purpose of this study is to determine the instrument's validity and reliability score, as well as to investigate the use of classical test theory to identify item characteristics. The research result in this study is the first step to build the instrument diagnostics based on Computerized Adaptive Test (CAT). The second step is to analyze the diagnostic instrument that produced by this research using the Item Response Theory.

Method

This type of study is descriptive exploratory research. The data are the responses of grade 5 and 6 elementary school students to the mathematics (Fractions, Comparison of two quantities, and Scale) diagnostic test instrument.

Research Design

Mathematics material in elementary schools has been determined through government regulations in Permendikbud Number 37 of 2018 for Elementary Schools and Madrasah Ibtidaiyah units. The regulation states that the curriculum objectives include four competencies, namely (1) spiritual attitude competencies, (2) social attitudes, (3) knowledge, and (4) skills. The diagnostic instrument developed in this study includes knowledge competence and skill competence described in the essential competencies of learning mathematics. As a result, the instrument for diagnosing elementary mathematics abilities (Fractions, Comparison of two quantities, and Scale) consists of knowledge and skill competence.

Population and Sample/ Study Group/Participants

The data collected were 166 respondents from students of 5 elementary schools in Cirebon, West Java, Indonesia, from November to December 2020.

Data Collection Tools

Five experts validated the instrument with a mathematics education background. The data analysis technique used is the analysis of the characteristics of the items based on the classical test theory using the CTT package on the R software and calculating content validity using the Aiken formula with five raters. Instrument reliability will be calculated using the Cronbach value and the Spearman-Brown formulation.

Data Collection

The number of items, as many as 20 items, is a material of diagnosing elementary school mathematical abilities (Fractions, Comparison of two quantities, and Scale). The elementary school mathematics diagnostic test instrument consists of 20 multiple choice questions with four alternative answer .

Data Analysis

Confirmatory Factor Analysis (CFA) is a form of factor analysis, especially in social research. The primary goal is to determine whether or not the indicators that have been grouped based on their latent variables (constructs) are consistent in the construct. In CFA, the researcher determined whether or not the data fit the previously formed model. The main distinction between CFA and Exploratory Factor Analysis (EFA) is that in CFA, researchers already assume that indicators belong to specific latent variables. In CFA, researchers created a hypothetical model based on a theoretical framework or previous research that was used as a guide. The diagnostic instrument developed in this study includes knowledge competence and skill competence described in the essential competencies of learning mathematics. As a result, the instrument for diagnosing elementary mathematics abilities consists of knowledge competence and skill competence. As a result, a construct model has been developed and will be tested. CFA tests it with Lisrel or Amos software. CFA is regarded as a subset of structural equation modeling (SEM).

The measures used in the CFA are the same as those used in SEM, namely the measure of the model's suitability with the data (fitness index). Chi-Square, RMSEA, GFI, and AGFI are examples of model suitability measures that will be used apart from the weight values of each indicator. One thing that is similar to these two statistical methods (EFA and CFA) is that they both use the variance of each manifest variable as a representation of the size of the contribution to the construct variable.

The CFA is used to validate the latent construct measurement model, in this case the instrument for diagnosing elementary school mathematical abilities. This instrument is made up of 20 items. The instrument is a dichotomous scale of "True" and "False." The measurement model's feasibility was tested in order to validate the existing measurement model.

Findings

For more than 90 years, the classical approach to test theory known as Classical Test Theory (CTT) has served as the foundation for educational and psychological measurement. This method addresses measurement error and test reliability, which is affected by individual test items. The R program's CTT package includes functions for item-level analysis and tests of dichotomous and polytomous response items. The following are the findings of the R program analysis (Table 1).

. The content validity of Aiken was calculated using the results of the assessment of the elementary school mathematics ability diagnostic instrument from five raters who worked as mathematics education lecturers and elementary school teacher education lecturers. The analysis of twenty diagnostic test items using Software R showed a validity value of 0.867 to 1, which means that each item has high validity. To facilitate the analysis of all items and constructs, the researcher label the factors as follows in (Table 2.

The following are the results of the CFA analysis using the SPSS AMOS 26 program.

In Figure 1, the 9th and 17th items do not have a significant relationship to the KP factor, while the 18th item has no significant relationship to the KK factor. As a result, The researcher excludes the three items from the instrument, so the results in Figure 2.

Figure 2 shows that all 17 items significantly relate to their respective factors/dimensions. Significance shows from the 3-star mark in the Plabel column. Using IBM SPSS AMOS 26 Software, the path diagram shows in Figure 3.

Figure 3 shows that the model fits the student response data, as shown in Table 3 below.

Table 3 shows that four of the eight test characteristics do not fit, and four fit the criteria. So it can be concluded that the last model analyzed using SPSS AMOS 26 meets the criteria. In other words, the data fits the model. As a result, the elementary school mathematics ability diagnostic instrument is valid.

Table 1: Aiken validity results using R Software

Item	*i1*	*i2*	*i3*	*i4*	*i5*	*i6*	*i7*	*i8*	*i9*	*i10*
V. Aikens	1	1	0,87	0,87	1	0,93	1	1	0,93	0,87
Item	*i11*	*i12*	*i13*	*i14*	*i15*	*i16*	*i17*	*i18*	*i19*	*i20*
V. Aikens	1	1	0,87	0,87	1	0,93	1	1	0,93	0,87

			Estimate	S.E.	C.R.	PLabel
i4	<---	KP	1,000			
i9	<---	KP	,194	,120	1,619	,105
i10	<---	KP	1,065	,138	7,712	***
i11	<---	KP	,726	,124	5,875	***
i12	<---	KP	,573	,116	4,963	***
i14	<---	KP	,585	,132	4,432	***
i15	<---	KP	,568	,127	4,478	***
i16	<---	KP	,647	,125	5,185	***
i17	<---	KP	,191	,110	1,743	,081
i20	<---	KK	1,000			
i19	<---	KK	1,605	,363	4,422	***
i18	<---	KK	-,290	,112	-2,583	,010
i13	<---	KK	1,526	,354	4,307	***
i8	<---	KK	1,226	,317	3,868	***
i7	<---	KK	,994	,258	3,848	***
i6	<---	KK	1,908	,409	4,668	***
i5	<---	KK	1,829	,400	4,568	***
i3	<---	KK	1,670	,377	4,433	***
i2	<---	KK	2,067	,433	4,775	***
i1	<---	KK	2,102	,435	4,834	***

Fig. 1: CFA results of 20 items

			Estimate	S.E.	C.R.	PLabel
i4	<---	KP	1,000			
i10	<---	KP	1,056	,136	7,738	***
i11	<---	KP	,718	,122	5,867	***
i12	<---	KP	,561	,114	4,895	***
i14	<---	KP	,587	,131	4,481	***
i15	<---	KP	,561	,126	4,454	***
i16	<---	KP	,639	,124	5,158	***
i20	<---	KK	1,000			
i19	<---	KK	1,616	,366	4,416	***
i13	<---	KK	1,532	,357	4,298	***
i8	<---	KK	1,232	,319	3,863	***
i7	<---	KK	1,005	,261	3,856	***
i6	<---	KK	1,925	,413	4,661	***
i5	<---	KK	1,831	,402	4,552	***
i3	<---	KK	1,671	,378	4,417	***
i2	<---	KK	2,070	,435	4,758	***
i1	<---	KK	2,098	,436	4,812	***

Fig. 2. CFA results of 17 items

The R program was used to perform classical test theory analysis. The total number of test items is 20, with 166 students taking the test. The syntax of the R program is as follows:

```
# Syntax in the R . program
install.packages('CTT')
library(CTT)

# student response data
dataD1 <- read.table('datadiagnosa.txt',header=T)
dataD <- dataD1[,-1]

# answer key
dataK <- c("D","B","D","A","C","B","C","C","B","D","C"
,"D","C","C","B","C","C","B","B","D");
dataK
```

Table 2: Factor Label

Factor	*Code*
Knowledge Competence	KP
Skill Competence	KK

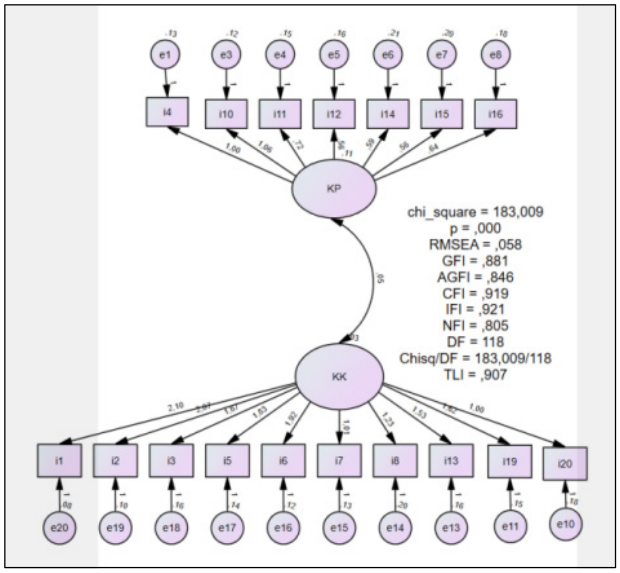

Fig. 3: Path Diagram of Diagnostic Instruments

```
# student score
score(dataD,dataK)

# score per item
myScores = score(dataD, dataK, output.scored=TRUE)
myScores$scored

#reliability with Cronbach value
items = myScores$scored
reliability(items)

# reliability with Spearman-Brown Formula
library(schoolmath)
subtest1=items[,is.odd(1:20)]
subtest2=items[,is.even(1:20)]
subscore1=apply(subtest1,1,sum)
subscore2=apply(subtest2,1,sum)
spearman.brown(cor(subscore1,subscore2),2,"n")

# item characteristics (difficulty level/p-value (itemMean),
discriminating power (pBis or bis))
itemAnalysis(items)
iA = itemAnalysis(items, itemReport=TRUE,
NA.Delete=TRUE, rBisML=FALSE)
iA$itemReport

# distractor answer choice analysis
distractorAnalysis(dataD,dataK)
```

Output Reliabilitas Cronbach :
Number of Items
20

Number of Examinees
166
Coefficient Alpha
0.88
Output Reliabilitas Spearman-Brown:
$r.new
[1] 0.889382

Table 3: Feasibility Test of Diagnostic Instruments

Name of Category	*Name of Index*	*Level of Acceptance*	*Analysis Results*	*Information*
Absolut Fit	Chi-Square	P Value > 0.05	0	Not Fit
	RMSEA	RMSEA < 0.08	0,058	Fit
	GFI	GFI > 0,90	0,881	Not Fit
	AGFI	AGFI > 0,90	0,846	Not Fit
Incremental Fit	CFI	CFI > 0,90	0,919	Fit
	IFI	IFI > 0,90	0,921	Fit
	NFI	NFI > 0,90	0,805	Not Fit
Parsimonious Fit	Chisq/df	Chi – Square /df < ,.0	1,55	Fit

The Cronbach coefficient is estimated to be 0.88, indicating that the internal consistency reliability for all 20 test items is satisfactory. The Alpha formula, according to Arikunto (2010: 239), is used to determine the reliability of an instrument whose score is not 1 or 0, such as a questionnaire or a question form description. The Spearman-Brown formula is used to determine the dependability of dichotomous or polytomous items. It divides the test into odd and even numbers (Azwar, 2012). The results of the Spearman-Brown formula were compared with the reliability coefficient categorization according to Guilford (1956: 145).

To calculate the reliability coefficient, used the R program with the result is 0.889 (the category of very high-reliability coefficient).

The p-value (itemMean) of all 20 items is between 0.2 and 0.8, which means that the items' difficulty level is moderate. At the 9th and 17th points, the biserial point correlation value is less than 0.2, which means that the discriminatory power for the two items is small (Sheng, 2019). In some references, the biserial point correlation value is also interpreted as the validity value of the item. The biserial point correlation value below 0.2 indicates that the item is not valid. As a result, items 9 and 17 are invalid and must be revised or not used. It follows the construct validity analysis using CFA. The biserial point correlation value shows that most questions have a discriminating index between 0.4 - 0.7 which is a "good" category (Arikunto, 2008).

Distractor Analysis Output (Table 5 and 6):

Here, the researcher only discusses the 9th and 17th items because they have a low discriminating index. In items 9 and 17, more students in mid50 (second quartile) answered correctly compared to mid75 (third quartile), and more high ability students (upper) answered at least one distractor than low ability students (lower). The discriminant value of the two items is also relatively consistent with the biserial point correlation (pBis). As a result, items 9 and 17 have low discriminatory power, so it is recommended that these items be revised. For item 9, distractors were chosen by about 68% of the test takers. As for item 17, distractors were chosen by about 76% of the test takers.

Table 4: Item Analysis Output:

	item Name	*itemMean*	*pBis*	*bis*	*alphaIfDeleted*
1	i1	0.6686747	0.6831713	0.8865060	0.8675412
2	i2	0.6204819	0.6573391	0.8380932	0.8682381
3	i3	0.5722892	0.5237147	0.6603537	0.8728267
4	i4	0.6084337	0.5719236	0.7267542	0.8711784
5	i5	0.5481928	0.5858084	0.7361615	0.8706434
6	i6	0.6204819	0.6142995	0.7832187	0.8697329
7	i7	0.7951807	0.4198361	0.5966444	0.8761646
8	i8	0.4156627	0.4151124	0.5245778	0.8765458
9	i9	0.3132530	0.1235567	0.1617070	0.8854634
10	i10	0.5843373	0.6373222	0.8053845	0.8688605
11	i11	0.6987952	0.5049883	0.6651290	0.8735202
12	i12	0.2530120	0.3971277	0.5398677	0.8768508
13	i13	0.6144578	0.5505476	0.7007282	0.8719202
14	i14	0.5000000	0.3769467	0.4724326	0.8778995
15	i15	0.6385542	0.3463962	0.4442091	0.8787301
16	i16	0.3253012	0.4566540	0.5941839	0.8750734
17	i17	0.2349398	0.1509912	0.2083411	0.8838395
18	i18	0.5481928	0.6487410	0.8152463	0.8684079
19	i19	0.6445783	0.6157243	0.7912794	0.8697419
20	i20	0.7048193	0.3776043	0.4990393	0.8775418

Table 5: Distractor Item i9

correct key	*n*	*rspP*	*pBis*	*discrim*	*lower*	*mid50*	*mid75*	*upper*
A	27	0.162	-0.360	-0.302	0.302	0.225	0.121	0.000
* B	52	0.313	0.123	0.291	0.232	0.250	0.243	0.5238
C	26	0.156	-0.297	-0.208	0.255	0.225	0.097	0.0476
D	61	0.367	0.111	0.219	0.209	0.300	0.536	0.4285

Table 6: Distractor Item i7

correct key	*n*	*rspP*	*pBis*	*discrim*	*lower*	*mid50*	*mid75*	*upper*
A	40	0.240	-0.292	-0.277	0.325	0.275	0.317	0.047
B	26	0.156	-0.190	-0.161	0.232	0.175	0.146	0.071
* C	39	0.234	0.150	0.311	0.093	0.250	0.195	0.404
D	61	0.367	-0.009	0.127	0.348	0.300	0.341	0.476

Discussion

Validity for measurement is closely related to the accuracy with which it is possible to measure an attribute using a test (Lovitt, 1993). Cronbach (1990) divides validity into two types, namely, logical validity–categories and empirical validity categories. Logical validity – categories for grouping approaches based on logical analysis, usually test content; and empirical validity – categories for clustering approaches based on empirical evidence, usually involving the correlation of test scores between tests.

Content validity is included in the logical validity type. Content validation typically involves a one-time study conducted by a panel of expert judges involved in a logical analysis of the relationship between test content and content specification for the domain (Lovitt, 1993). Content validity was analyzed using the Aiken formula. The data for analysis uses the assessment results of the elementary school mathematics ability diagnostic instrument from five raters who work as mathematics education lecturers and Elementary School Teacher Education Lecturer. The analysis of twenty diagnostic test items using Software R showed a validity value of 0.867 to 1, which means that each item has high validity. However, content validity alone is not enough. Logical and empirical evidence about the quality of judgments is a necessary component of validation (Lovitt, 1993).

Empirical evidence has been analyzed using construct validity. According to Messick (Lovitt, 1993), proper validation - namely, construct validation - is concerned with tests and the justification of interpretation of test responses and scores. Construct validity was analyzed using Confirmatory Factor Analysis. The analysis showed that 3 out of 20 items had no significant relationship with the factors/dimensions, so the three items were excluded from the diagnostic instrument. It has been shown that the diagnostic instrument containing 17 items has a model fit with the data, so the instrument is valid.

After the instrument is proven valid, the next step is to prove its instrument's reliability. The Cronbach coefficient is approximately 0.88. It shows that the internal consistency reliability for all 20 test items is at least 0.88, indicating that it is satisfactory. The Alpha formula, according to Arikunto (2010: 239), is used to determine the reliability of an instrument whose score is not 1 or 0, such as a questionnaire or a question form description. Because the diagnostic instrument in this study is dichotomous data, the reliability calculation using the Cronbach coefficient will be biased. As a result, researchers use the Spearman-Brown formula to investigate the score reliability. After the instrument's validity and score reliability are met, the item characteristics will be analyzed based on the classical test theory.

According to classical test theory, the apparent score (X) is made up of the true score (T) and the error score (E). The fundamental assumption in classical test theory is that there is no correlation between the actual and error scores, and that the mean random measurement error is zero (Allen & Yen, 1979). Some formulas for calculating the test reliability index were developed based on these assumptions.

Mardapi (1998) defines three estimated item parameters: level of difficulty, discriminating power, and level of guesswork. The difficulty level is the proportion of participants who incorrectly answered the items, which can be calculated by comparing the number of test participants who correctly answered the items to the number of test participants who incorrectly answered the items. The discriminating power for each item then provides information about items that can differentiate test takers' abilities. Discriminatory power is defined as the correlation between test item scores and total scores, also known as biserial point correlation.

Ebel and Fresbie (1986:234) define biserial point correlation as the relationship between test item scores and total score for each testi. Biserial with positive and high scores indicates that testees with high scores are more likely to correctly answer, whereas testees with low scores are more likely to incorrectly answer. Negative biserial scores, on the other hand, reveal information about high-scoring testers' proclivity to provide incorrect answers when answering items. In contrast, low-scoring testies correctly answer these items. Items with negative biserial points should be excluded from the model during item analysis to select good items. In the diagnostic instrument developed in this study, all items' biserial point correlation value is positive. The biserial point correlation value below 0.2 indicates that the item is not valid. As a result, items 9 and 17 are invalid and must be revised or not used. It follows the construct validity analysis using CFA.

Sudijono (2009: 376-378) suggests that for items whose difficulty index is sufficient or moderate, it should be immediately recorded in the question bank. For items categorized as easy, the items are re-examined to find out the cause. According to the Ministry of National Education (2010:11), the cause could be that the item's distractors did not work, or it could be that the majority of students answered the item correctly, indicating that the majority of students understood the material being tested. According to the Ministry of National Education (2010:11), a difficult item could be because the answer key is incorrect. It has two or more correct answers. It has never been taught, or the learning is unfinished. Maybe it is not suitable to be asked using the form of the question, or the question sentence is too complex and lengthy.

In terms of discriminatory power, Sudijono (2009: 408-409) suggests that items with good distinguishing power (satisfactory, good, and excellent) should exist in the question bank. For items with low discriminatory power (poor), the tester (teacher) should trace them and correct them. After being

corrected, it can be submitted again in the upcoming learning outcomes test; later, the item is analyzed again, whether the discriminatory power increases or not.

For items with a negative index of item discrimination, it is better not to use them anymore. According to the Ministry of National Education (2010:11), if an item cannot distinguish between two abilities of students, the following scenarios may occur in the question: the answer key to the item is incorrect; the item has two or more correct answer keys; competence measured is not apparent; the distractor does not work; the material asked is too tricky, so many students guess; or the majority of students who understand the material being asked believe there is something wrong with the item.

Furthermore, the thing that must be considered in making multiple choice questions is the use of distractors. According to Arikunto (2008: 220), a distractor can function well if it has an excellent appeal for test takers who do not understand the concept. Meanwhile, according to Sudijono (2009: 411), distractors can perform their functions properly if at least 5% of all test takers have been selected. This study's instrument for diagnosing elementary school mathematics abilities had 20 items. Items with distractors below 5% are item number 5 option D (1.8%), item number 7 option A (1.8%), item number 8 option D (1.8%), item number 18 option C (4.8%), and item number 19 option D (1.8%). Of the 60 total distractors, 5 (8.3%) did not function well because they were chosen by less than 5% of the participants. Meanwhile, 55 distractors (91.7%) functioned well because they were chosen by more than 5% of the test takers. Based on the explanation above, the analysis of validity, reliability, and item characteristics helps produce a good instrument that can be used practically.

Conclusion

According to the findings of the content validity analysis, the mathematics ability diagnosis instrument had high validity ranging from 0.87 to 1, and the reliability score was 0.889. While the results of the analysis of construct validity using CFA obtained information that there are three items, namely the 9th, 17th, and 18th items, which are not significantly related to the factors/dimensions, the three items are excluded from the model/instrument. The CFA approach also informs that the instrument for diagnosing elementary mathematics abilities, consisting of 17 items, is valid. Based on the index of difficulty level, the diagnostic instrument developed, as a whole, is categorized as a moderate item because it has a value above 0.2 and below 0.8. There are two items, namely items 9 and 17, with low discriminating power, so it is recommended that the two items be revised or not used. Most questions have a discriminatory index value between 0.4 - 0.7, categorized as good. Of the 60 total distractors, 5 (8.3%) did not function well because they were chosen by less than 5% of the participants. Meanwhile, 55 distractors (91.7%) functioned well because they were chosen by more than 5% of the test takers. The diagnostic instrument produced from this study can be used to diagnose elementary school mathematics abilities for mathematics lessons up to 5th-grade material.

Suggestion

Using classical test theory, this study provides an overview of determining the quality of the items that comprise the instrument. These results can be used primarily by practitioners in determining the quality of the instrument before it is used to diagnose students' mathematical abilities. Furthermore, this analysis is also expected to be used to analyze the quality of the instrument measuring the achievement of student learning outcomes. To strengthen the analytical evidence, the researcher should continue with analysis using modern test theory, involving more sample sizes.

Limitation

This study has several limitations. First, the test instrument that was analyzed only covered 5th-grade mathematics. Second, classical test theory is used because it considers sample size and the feasibility of the analysis. Third, data collection was carried out only on students in five elementary schools by considering the ease of access in the study.

Acknowledgments

Thanks to the Government of Indonesia for funding this research through the Directorate of Resources, Directorate General of Education, Ministry of Education, Research and Technology, in accordance with the Funding and Research Contract of 2022 Number: 127/E5/PG.02.00.PT/2022.

References

Allen, M.J., & Yen, W.M. (1979). Introduction to Measurement Theory. California: Wadsworth, Inc.

Arikunto, Suharsimi. 2008. Dasar-Dasar Evaluasi Pendidikan. Jakarta: Bumi Aksara.

Arikunto, S. (2010). Prosedur Penelitian Suatu Pendekatan Praktek. Jakarta: Rineka Cipta.

Andrich, D., & Marais, I. (2019). A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences (Issue 1989, pp. 41–53). https://doi.org/10.1007/978-981-13-7496-8

Bodner, G. M. (1986). Constructivism: A theory of knowledge. Journal of Chemical Education, 63(10), 873–878. https://doi.org/10.1021/ed063p873

Cronbach, L. J. (1990). Essentials of Psychological Testing. Harper & Row.

Hambleton, R. K. (2005). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. Educational Measurement: Issues and Practice, 12(3), 38–47.

Istiyono, E., Mardapi, D., & Suparno, S. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta Didik SMA. Jurnal Penelitian Dan Evaluasi Pendidikan, 18(1), 1–12. https://doi.org/10.21831/pep.v18i1.2120

Kurian, G. (2014). Reliability and Validity Assessment. In The Encyclopedia of Political Science. https://doi.org/10.4135/9781608712434.n1341

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). Cognitive Diagnostic Assessment for Education: Theory and Applications. Cambridge University Press.

Levine, M. D., Lindsay, R. L., & Reed, M. S. (1992). The wrath of math: Deficiencies of mathematical mastery in the school child. Pediatric Clinics of North America, 39(3), 525–536. https://doi.org/10.1016/S0031-3955(16)38342-0

Lovitt, R. (1993). Psychological Assessment. In Journal of Personality Assessment (Vol. 60, Issue 3). https://doi.org/10.1207/s15327752jpa6003_20

Retnawati, H. (2016). Analisis Kuantitatif Instrumen Penelitian (Pertama). Parama Publishing.

Sharkness, J., & DeAngelo, L. (2011). Measuring Student Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys. Research in Higher Education, 52(5), 480–507. https://doi.org/10.1007/s11162-010-9202-3

Sheng, Y. (2019). CTT Package in R. Measurement, 17(4), 211–219. https://doi.org/10.1080/15366367.2019.1600839

Tall, D., & Razali, M. R. (1993). Diagnosing students' difficulties in learning mathematics. International Journal of Mathematical Education in Science and Technology, 24(2), 209–222. https://doi.org/10.1080/0020739930240206

Traub, R. E. (2005). Classical Test Theory in Historical Perspective. Educational Measurement: Issues and Practice, 16(4), 8–14.

Zimmerman, D. W. (1998). How should classical test theory have defined validity? Social Indicators Research, 45(1–3), 233–251. https://doi.org/10.1023/a:1006949915525