Eurasian Journal of Educational Research
www.ejer.com.tr

Using Alignment Index and Polytomous Item Response Theory on Statistics Essay Test

Kana HIDAYATI[1], BUDIYONO[2], SUGIMAN[3]

ARTICLE INFO

Article History:
Received: 2 Apr. 2018
Received in revised form: 10 Oct. 2018
Accepted: 23 Dec. 2018
DOI: 10.14689/ejer.2019.79.6

Keywords
Alignment, validity, reliability, IRT polytomous GPCM, essay test

ABSTRACT

Purpose: Essay test in mathematics, both in the form of restricted-response and extended-response, generally consist of polytomous scored items. However, the essay test used by teachers in Indonesia has not been fully supported by sufficient quality evidence. There have been many studies focusing on the development of the essay test, but not many of them have applied the use of relevant measurement theory for the polytomous data. The evidence of content validity also has not been supported by its alignment with the curriculum. This study used alignment index to prove the content validity and IRT polytomous GPCM to determine the characteristics of test items in order to produce an essay test that could accurately measure the achievement of students on statistical materials.
Method: Procedures of this study: (1) preparation of preliminary test, (2) trials, (3) interpretation. Trial was conducted involving 688 Junior High School students in Yogyakarta, Indonesia.
Results: The content validity of the test was good, supported by V Aiken index of 0.88–1.00 and Porter alignment index of 0.93. The test items had good construct validity. Test reliability was categorized as good with the Construct Reliability coefficient of 0.88 and the Alpha coefficient of 0.78. Judging from its characteristics, all test items were categorized as good.
Implications for Research and Practice: The use of the alignment index contribution to the verification of content validity of essay test and the use of the IRT polytomous GPCM may provide reference for the use of appropriate measurement theory to determine the item characteristics of essay test.

[1] Corresponding Author: Yogyakarta State University, INDONESIA, e-mail: kana@uny.ac.id, ORCID: https://orcid.org/0000-0002-9226-8500
[2] Sebelas Maret University, INDONESIA, e-mail: budiyono@staff.uns.ac.id, ORCID: https://orcid.org/0000-0001-6467-7801
[3] Yogyakarta State University, INDONESIA, e-mail: sugiman@uny.ac.id, ORCID: https://orcid.org/0000-0002-9226-8500

Introduction

Assessment is an important component for the successful achievement of learning mathematics in school. Some of the things assessed in the learning of mathematics include the understanding of concepts and problem-solving skills. Based on Curriculum 2013 in Indonesia, this is manifested in the form of an assessment of students' mathematical knowledge competence achievement. The use of the instrument in the form of essay test in the assessment of mathematical knowledge, especially on statistical material is very beneficial to be selected by teachers.

Essay tests in mathematics can be presented in various formats. In general, essay formats are usually classified into two groups: restricted response items and extended-response items (Nitko & Brookhart, 2011, p. 204). The development of mathematics learning shows that both types of essay tests can be used in mathematical assessment. The advantages of the essay test by Walstad (2006, p. 4) are: (1) it has a great potential to assess students' level of understanding at a higher level, (2) students have the freedom to prepare, choose, and present ideas in their own words while answering, (3) teachers have the opportunity to see their students answers, (4) it is suitable for achievement tests related to problem analysis, concept application, or decision evaluation. Therefore, the essay test is well suited to measure and assess the achievement of students' mathematical knowledge competence.

The use of a valid and reliable instrument that meets the criteria as a good item will provide an accurate and accountable assessment result. Validity evidence of an instrument generally includes content validity and construct validity. Evidence of content validity is done by rational analysis through expert judgment and evidence of construct validity is provided by factor analysis. Good assessment instruments, in addition to validity, should also be reliable. Reliable assessment instruments will give relatively the same results on each measurement, although measurement times are different.

Related to the validity of the content, it is usually supported by the calculation of the content validity index. One of the approaches to determine the coefficient of content validity is proposed by Aiken (1985, p. 132). Formula of V Aiken to calculate the index of content validity is based on the result of the assessment of several experts against an item in terms of the extent to how much the item represents the measured domain or contsruct. However, the development of measurement theory shows that the validity of the contents of an assessment instrument can also be obtained through alignment tests between assessments with standards in the curriculum. Ananda (2003) and Bhola, Impara, and Buckendahl (2003) stated that the alignment test results can be used as evidence of content validity. Until now, there are very little informations about the alignment test related to the assessment with the standards set by the government in the curriculum in Indonesia, especially in learning mathematics.

In Indonesia, various statistics' essay tests developed by researchers in general have been supported by evidence of the quality of the instrument related to content validity through expert judgments such as the ones developed by Buhaerah (2010), Hanjarwati and Wiyarno (2015), and Effendi and Farlina (2017). However, not many of various studies that contain statistics' essay test in Indonesia nowadays have been supported by the evidence of content validity using index alignment. Many of them also have not been supported by evidence of construct validity and item characteristics by using relevant measurement theory such as Item Respons Theory (IRT) for the polytomous data. Application of measurement theory in statistics tests have been conducted by Guler (2014) who analyzed open ended statistics questions with many facets of Rasch model. The use of IRT polytomous to date has not been widely applied, especially in the essay test. In fact, the assessment instrument in the form of essay test in mathematics learning especially related to statistical materials is generally arranged using response format in more than two categories (polytomous). One suitable model of IRT polytomous that is used in scoring the item test response is Generalized Partial Credit Model (GPCM). The assumption in GPCM usage is that the test items have different levels of difficulty and the level of difficulty of each step is not sorted out. This is quite relevant to the assessment generally done by teachers in Indonesia by providing an essay test score that is based on the number of steps answered correctly without regard to the sequence of steps.

Evidence of Content Validity with Alignment Index

Content validity aims at exploring whether the contents of a measuring instrument is representative or not in order to measure intended performance domain (Crocker & Algina, 1986, p. 218). Sireci and Bond (2014, p. 100) state that the evidence of the validity of the contents of an instrument, especially tests can be conducted through traditional and modern approaches. Traditionally, the most common method used to prove content-based validity is through expert judgment. The evidence of content validity is supported by the magnitude of the content validity coefficient of Aiken (1985, p. 132). As for the modern, new developed approach related to the validity of the content is conducted through the test of alignment between assessment and standards. Biggs (2003, p. 14) states that it is difficult to accurately obtain student achievement information in accordance with the desired objectives when the assessment is not in accordance with the standards in the curriculum. Furthermore, according to Wiggins and McTighe (2001, p. 51), without such conformity it limits the achievement of the expected outcomes because the students will not be studying what is being assessed.

Some of the current alignment methods include: (a) Webb Method, (b) La Marca Method, (c) Survey of Enacted Curriculum (SEC) Method, (d) Bloom's revised taxonomy Method, and (e) Method of alignment Project 2061 from the American Association for the Advancement of Science (AAAS). Empirical studies show that Bloom's revised taxonomy can be used as a tool for aligning test results in a higher level of inter-rater reliability than some other taxonomies (Nasstrom & Henriksson, 2008). Developments in alignment studies indicate that Bloom's revised taxonomy method for testing alignment between assessment and standards in the curriculum can be modified with the Porter model in terms of calculating its alignment index. Alignment index ranges from 0 (no alignment) to 1 (perfect alignment). Nasstrom and Henriksson (2008) conducted a study of alignment between assessment and standards in the curriculum by using Bloom's revised taxonomy of the associated cognitive complexity. The formula of alignment index of Porter (*P*) is as follows.

$$P = 1 - \frac{\sum_{k=1}^{k}\sum_{j=1}^{j}|a_{jk}-b_{jk}|}{2}$$

Where: *j* is the number of rows, *k* is the number of columns in each matrix *X* and *Y*, a_{jk} and b_{jk} is the ratio in cells in row *j* and column *k* for each *x* and *y* ratio matrix. Research on alignment studies by Nasstrom and Henriksson (2008) in Sweden shows that Bloom's revised taxonomy is the best model to prove harmony especially in mathematics subjects.

Item Response Theory Polytomous GPCM

Several models of the proposed item response theory polytomous are: Nominal Response Model (NRM), response model for multiple-choice items, Rating Scale Model (RSM), Partial Credit Model (PCM), Graded Response Model (GRM), sequential model for ordered response, and the Generalized Partial Credit Model (GPCM) (Van der Linden & Hambleton, 1997, p. 30). Thorpe and Favia (2012) stated that the assumptions that must be met in the analysis of test items based on IRT polytomous are sample size and unidimensionality of data. The sample size in IRT polytomous according to Reeve and Fayers (2005) is at least 250, but a sample size of about 500 is preferable for the accuracy of parameter estimation.

GPCM is one of the suitable models used to learn the characteristics of test items used in Indonesia. This is because the math test items in Indonesia are generally scored using a partial credit system that is the answer to each settlement step to the right answer is appreciated and the level of difficulty of each step is not sequenced. The general form of GPCM is stated as follows (Muraki, 1993, p. 351-352).

$$P_{jk}(\theta) = \frac{exp\left[\sum_{v=1}^{k} Z_{jv}(\theta)\right]}{\sum_{e=1}^{m_j} exp\left[\sum_{v=1}^{e} Z_{jv}(\theta)\right]}$$

$$\text{and} \quad Z_{jv}(\theta) = Da_j(\theta - b_{jv}) = Da_j(\theta - b_j + d_v)$$

Where: $P_{jk}(\theta)$ is the probability of a participant with ability θ who obtains k score category on item j, D is a scaling constant that puts the trait (θ) scale in the same metric as the normal ogive model (D=1.7), a_j is a slope parameter, b_{jh} is an item-category parameter, b_j is an item location parameter, d_v is a category parameter, m_j+1 is number of item in j, and *D* is the scale factor (D=1.7). Estimation of item parameters and ability on IRT politomus can be done with the help of Parscale software from SSi (Muraki & Bock, 1997).

Based on the item response theory, the function of the item information states the strength or contribution of the item in uncovering the latent trait measured by the test. The function of the item information on the item response of polytomous is given by Samejima (Muraki, 1993) as follows.

$$I_j(\theta) = D^2 a_j^2 \sum_{c=1}^{m_j} \left| T_c - \bar{T}_j(\theta) \right|^2 P_{jc}(\theta)$$

Where: $\bar{T}_j(\theta) = \sum_{c=1}^{m_j} T_c P_{jc}(\theta)$, $I_j(\theta)$ is the information function of item j, D is a constanta which can have value of 1 or 1.7, a_j is a slope parameter of item j. $\bar{T}_j(\theta)$ item response function for a polytomous-scored item. Based on the value of the function, the item information can be determined by the function value of the test information ($I(\theta)$) and the estimated value of Standard Error Measurement ($SEM(\hat{\theta})$) with the following formula (Hambleton, Swaminathan, & Rogers, 1991, p. 94).

$$I(\theta) = \sum_{j=1}^{n} I_j(\theta) \quad \text{and} \quad SEM(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$$

This study is using alignment index and IRT polytomous GPCM in order to make sure that essay test produced in this study can accurately measure the achievement of students' knowledge competencies on Junior High School statistical materials. Theoretically, the test of quality includes content validity through expert judgment supplemented by V Aiken's index calculation and alignment index. The empirical evidence includes: (a) testing of construct validity using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA), (b) determining item characteristics, information function, and standard error measurement using IRT polytomous GPCM approach; and (c) determining Construct Reliability coefficient and reliability the test by its internal consistency with Cronbach's Alpha.

Method

Research Design

The present study used the development research and descriptive survey. In this study, the alignment index and polytomous item response theory were used in order to analyze the answers given to nine items statistics' essay test.

Research Sample

Participants in the trials were 688 Junior High School students of class VII in Yogyakarta city of Indonesia that applied Curriculum 2013. Selection of school was done by purposive cluster sampling technique which was based on the category of school and certain considerations. The tests in this study were general and aimed at all students with low, moderate or high ability. Therefore, trials were conducted on three junior high schools representing high, medium, and growing qualities as well as representing public and private schools. Clasification of school quality was done based on the results of the National Examination of 2015/2016 academic year lessons.

The selected school were SMP N 5 Yogyakarta which represented high quality, SMP IT Abu Bakar Yogyakarta which represented medium quality, and SMP Muhammadiyah 2 Yogyakarta which represented growing quality.

Procedures

The arrangement of statistics' essay test in this study used modification of the instrument development model from Wilson (2005, p. 18-19) and the development of test instruments from Oriondo and Antonio (1998, p. 34). Procedures of this study involved three steps namely (1) preparation of preliminary test, (2) trials, and (3) interpretation of trial results. Preliminary tests were designed by referring to the basic competencies, indicators, and contents of Bloom's revised taxonomy to each item as presented in Table 1.

Table 1

Basic Competencies, Indicators, and Bloom's Revised Taxonomy

Basic Competencies	*Indicators*	*Item*	*Bloom's Revised Taxonomy*
Analyze the relationship between data and the way of presentation (table, line graph, bar chart, and pie chart).	1. Describes various ways of collecting data.	1a	B3
	2. Describes various ways of presenting data.	1b	B3
	3. Presents data using tables.	2a	C3
	4. Analyzes the relationship between the data presented in tabular form.	2b	C4
	5. Presents data using bar charts.	3a	C3
	6. Presents data using pie charts.	3b	C3
	7. Analyzes the relationship between the data presented in the form of bar and circle diagrams.	3c	C4
	8. Presents data using line graphs.	4a	C3
	9. Analyzes the relationship between the data presented in the form of a line graph.	4b	C5

Description: B3: Conceptual knowledge and cognitive processes "apply", C3: Procedural knowledge and cognitive processes "apply", C4: Procedural knowledge and cognitive processes "analyze", C5: Procedural knowledge and cognitive processes "evaluate".

Data Analysis

Content Validity

Test validation was performed through expert judgment. Experts conducted a qualitative test review and provided an assessment of the suitability between the item with the indicator in the form of a Likert scale with five answer options. In addition, to strengthen the evidence of content validity for the purpose of alignment tests, the review sheet also featured a format for the assessment of the expert on the revised content of Bloom's taxonomy on each test item. Based on the assessment of experts, in addition to qualitative assessment related to feasibility of the test, V Aiken's index was also calculated. The formula of V Aiken is as follow:

$$V = \frac{s}{[n\,(c-1)]}\,,\ s = \Sigma\, n_i\, (r_i\text{-}l_o)$$

Where: V is Aiken validity index, n_i is the number of experts who choose the criteria of i, r_i *is* the criteria of i, l_o is the lowest rating, n is the number of expert, and c is the number of rating.

The valid criterion of an item is to compare the value of V calculated with V value, that is the minimum value of the content validity index based on the number of rater in V table Aiken (1985). The number of raters in this study was six and the number of ratings was five then the minimum index of content validity based on table V Aiken was 0.79. The result of the expert judgment was a proof of theoretical quality of the instrument. The empirical evidence was obtained from the trial and interpretation of test results.

Construct Validity

To obtain evidence of construct validity, factor analysis using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) were used. Using the EFA approach, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO MSA) and Barlett Sphericity tests were used to determine whether the items' test matched the factor analysis or not. The match criteria were a minimum KMO MSA score of 0.50 and statistically significant Barlett Sphericity test results (Hair et al., 2010). This study extracted factors using Principal Component Analysis (PCA) conducted using SPSS version 20.0. Using CFA approach, data analysis began with the requirements analysis test in order to determine whether the data already met the requirements to be analyzed by CFA technique, which required to test the model by using the joint multivariate normal distribution. In this study, CFA was conducted with Structural Equation Modeling (SEM) by using Lisrel program version 8.51. The criteria used were if p value was >0.05; then, the distribution is normal, and if p value was ≤0.05; then, the distribution is not normal (Yamin & Kurniawan, 2009).

After the requirements analysis test, data analysis was performed to verify the validity of scale constructs through first order CFA. The criteria for a valid indicator in representing the construct were if t value was>1.96 and the value of Standarized Loading Factor (SLF) was at least 0.3 (Hair et al., 2010; Igbaria et al., 1997). The criteria of sample size in SEM according to Comrey and Lee (1992) cited by MacCallum et al. (1999, p. 840) are: 100 (poor), 200 (fair), 300 (good), 500 (very good), and ≥1000 (excellent). This study involved 688 participants; thus, the sample size of the study belonged to the very good category.

Based on the data of the test results, the validity of the construct was conducted through the first order Confirmatory Factor Analysis (CFA) with the help of Lisrel software. The criterion of the validity of an item in representing the construct was the value of t value>1.96 and the value of Standarized Loading Factor (SLF) of at least 0.3 (Hair et al. 2009, p. 119; Igbaria et al., 1997, p. 290). The model fit criteria used were

Root Mean Square Error of Approcimation (RMSEA) between 0.03 to 0.08, p value>0.05, Goodness of Fit Index (GFI)≥0.90, Adjusted Goodness of Fit Index (AGFI)≥0.90 (Hair et al, 2010, p. 641-644).

Those criteria are based on Garson (2009) which state that support for the fit of the model developed through empirical data can be seen at least from three compatibility measures representing three different fit model categories. The three categories of fit model are absolute fit measures, incremental fit measures, and parsimonious fit measures. If two of the three categories meet the criteria, then the developed model matches the data. Therefore, the criteria for model fit used in this study were RMSEA, p value, GFI which represented absolute fit measures, and AGFI which represented incremental fit measures. The fulfillment of unidimensionality assumptions was also seen from the plot of eigenvalue (Hambleton, Swaminathan, & Rogers, 1991, p. 56). Naga (1992, p. 297) stated if eigenvalue of the first factor several times the eigenvalue of the second factor, while the eigenvalue of the second factor and above are almost the same, it can be said that the unidimensional requirement has been fulfilled.

Characteristics of Test Items

Characteristics of test items were obtained through item analysis based on IRT polytomous GPCM with the help of Parscale version 4.1. Based on the results of Parscale analysis, three characteristics items were obtained; namely, the estimation of discrimination parameters (*a*), location parameters (*b-global*), and a set m-1 parameter of difficulty level (*b*). The b_{jk} parameter was obtained by subtracting the parameter value b-global$_j$ with the parameter value d_{jk}. The resulting graphics included Item Characteristic Curve (ICC), Item Information Curve (IIC), and Test Information Curve (TIC). The good item criteria were: (1) the parameter values *a*>0.25 on the logit scale, and (2) *b-global* parameter values ranged from -3 to 3 on the logit scale (Wells, Hambleton, & Purwono, 2008).

Reliability

One popular method in psychometry which is often used to determine reliability based on internal consistency is the alpha coefficient from Cronbach. Coefficient alpha can provide a reliability estimate for a measure composed of items scored with values other than 0 and 1 (Cronbach, 1951). The formula for estimating the reliability of essay test scores uses the basic coefficient alpha is as follows (Ebel & Frisbie, 1991, p. 85).

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_t^2}\right)$$

Where: k is the number of separately scored essay test questions, s_i^2 is the variance of students'scores on a particular item, $\sum s_i^2$ is the sum of the item variances for all test items, and s_t^2 is the variance of the total essay scores.

A good reliability criterion is a minimum of 0.7 (Nunnally, 1981, p. 245). As according to Ebel & Frisbie (1991, p. 86) if the test is used as a standard test, the reliability coefficient should be between 0.85-0.95 while for the minimum class, it is not lower than 0.65. Kayapınar (2014, p. 114) stated that reliability coefficient value might be more accurate and reliable if the accepted interpretation of a meaningful correlation coefficient for this kind of measurements can be considered as .90 minimum for giving evidence of reliable ratings. In addition to Alpha's Cronbach, on the use of SEM can also be obtained Construct Reliability (CR). The formula for calculating CR is as follows (Wijanto, 2008).

$$Construct\ Reliability = \frac{(\sum std.loading)^2}{(\sum std.loading)^2 + \sum e_j}$$

Where: std. Loading is Standarized Loading Factor (SLF), e is error variances. Hair, et al (2010) suggests that the estimation of CR>0.7 is good, while CR between 0.6 and 0.7 is acceptable, provided that the construct validity indicator is good.

Results

The validity of test content was conducted through qualitative and quantitative analyses. Qualitative results were statements by experts who claimed that the test was feasible and ready for use. Quantitative results included the assessment of experts on the suitability of the item with the indicator and the revised charge of Bloom's taxonomy. Based on the result of calculation of V Aiken's index as presented in Table 2, it was found that V of test items were 0.88-1.00. This meant that all test items had good content validity in terms of their conformity with the indicator.

Table 2

V Aiken's Index Calculation Result

Item	*s*	*n*	*c-1*	V_{table}	*V*
1a	21	6	4	0.79	0.88
1b	21	6	4	0.79	0.88
2a	24	6	4	0.79	1.00
2b	24	6	4	0.79	1.00
3a	22	6	4	0.79	0.92
3b	22	6	4	0.79	0.92
3c	22	6	4	0.79	0.92
4a	23	6	4	0.79	0.96
4b	23	6	4	0.79	0.96

Evidence of the validity of this content was also supported by the calculation of Porter alignment index that was equal to 0.93. The magnitude of the alignment index was included in the category of excellent so it can be said that the test had a very good alignment with the standards set out in Curriculum 2013 in Indonesia. The test that was declared as eligible by the experts was then tested on the trial with 688 students.

Based on trial, EFA was conducted in order to identify the factors that made up the test. Based on SPSS version 20.0, the value of Barlett Test of Sphericity was 3105.039 with 0.000 significance level. This showed a significant correlation between variables. The calculation result of KMO MSA was 0.729 which indicated that the adequacy of the sample was good. Factor extraction was done using PCA method. Based on the extraction results, three factors forming the test constructs were obtained. The findings related to the factors are given in Table 3.

Table 3

Findings Related to Factors Obtained as a Result of the Principal Component Analysis

Factor	*Eigenvalue*	*Variance Percentage*	*Total Variance Percentage*
1	3.745	41.606	41.606
2	1.684	18.713	60.319
3	1.062	11.801	72.120

Based on Table 3, it can be seen that there were three factors with an eigenvalue bigger than 1.00. All three factors explained around 72.120% of the total variance. The first factor described 41.606% of the total variance. The contribution of the factors to the total variance percentage decreased after the first factor. Eigenvalue of the first factor was 3.745, more than twice from eigenvalue of the second factor. That means the test only measured one ability. This situation can be seen in eigenvalue graph on Figure 1.

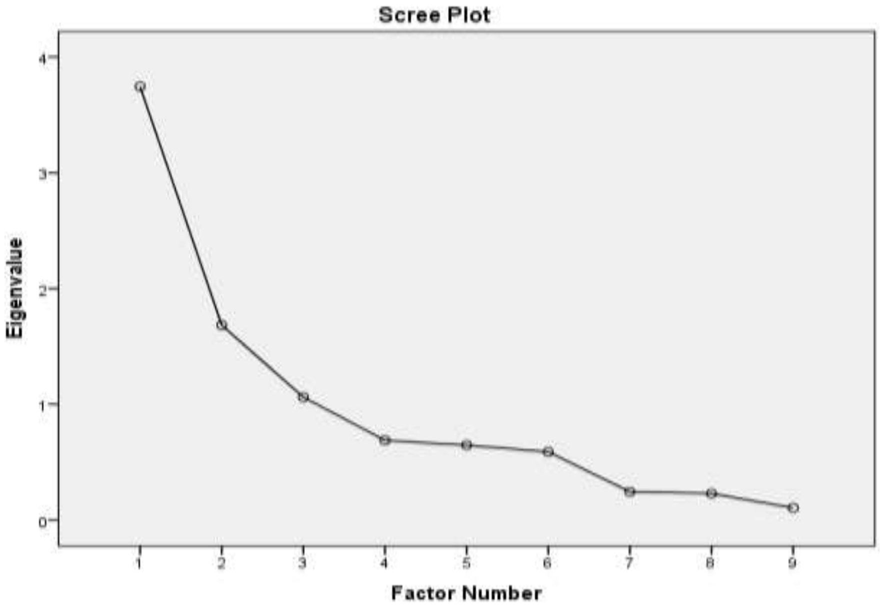

Figure 1. Eigenvalue Graph

For the next step, CFA was conducted in order to confirm that items really explained the test. The result of preliminary analysis showed that the data did not have multivariate normal distribution, so CFA second order analysis was done with Weighted Least Square (WLS) estimation model. Based on Lisrel analysis results, RMSEA value of 0.031, Chi Square of 20.11 with p value 0.06, GFI of 0.99, and AGFI of 0.98 were obtained. These results indicated that the suitability of the model was met. Here were CFA's fisrt order results for t value and Standarized Loading Factor (SLF) values as well as CR and Alpha coefficients.

As shown on Table 4, all of the items were significant in supporting the test constructs with the lowest support by item 1a and the highest by item 4a. It showed that the construct validity of the test was good. Test reliability was good with a CR coefficient of 0.88 and an Alpha coefficient of 0.78. This meant that in terms of construct, the test consisted of items that could accurately measure students' statistical skills. Based on the reliability scores, it can be said that the test was reliable which meant that the measurement results obtained through this test were consistent.

The characteristics of the test items were determined using the IRT polytomous GPCM approach. For the analysis of the data using Parscale software version 4.1., the following parameters were obtained.

Table 4

Result of First Order CFA

Item	*First order CFA*				*Validity*	*Reliability*	
	t value	*Explanation*	*SLF*	*Error*		*CR*	*Alpha*
1a	5.80	Significant	0.24	0.18	Sufficient	0.88	0.78
1b	7.46	Significant	0.48	0.19	Good		
2a	22.8	Significant	0.89	0.099	Good		
2b	7.78	Significant	0.33	0.54	Good		
3a	6.26	Significant	0.48	0.12	Good		
3b	14.74	Significant	0.77	0.88	Good		
3c	9.34	Significant	0.37	0.53	Good		
4a	24.38	Significant	0.99	0.0067	Good		
4b	16.58	Significant	0.44	0.84	Good		

As shown in Table 5, all the test items were categorized as good with the discrimination index (a_j) as a whole located at 0.310-2.008 and with the item difficulty index (b_j) of -2.329 to 0.475 on the logit scale. The analysis also obtained the function of test information and standard error measurement as presented in Figure 2.

Table 5

Results of Parameter Test Estimation on Trial

Item	a_j	b_j	d_k	b_{jk}	*Explanation*
1a	1.216	-1.849	0.000	-1.849	Good
1b	0.905	-2.179	0.183 -0.183	-2.362 -1.996	Good
2a	1.371	0.475	0.716 -0.716	-0.241 1.191	Good
2b	0.865	-1.086	0.171 -0.171	-1.257 -0.915	Good
3a	2.008	-2.329	-0.059 0.059	-2.270 -2.388	Good
3b	0.359	-1.691	-3.745 3.662 1.981 -3.214 1.316	2.054 -5.353 -3.672 1.523 -3.007	Good
3c	0.813	-0.977	0.008 -0.008	-0.985 -0.969	Good
4a	1.314	-0.715	-0.353 0.353	-0.362 -1.068	Good
4b	0.310	0.010	0.575 0.545 -1.121	-0.565 -0.535 1.131	Good

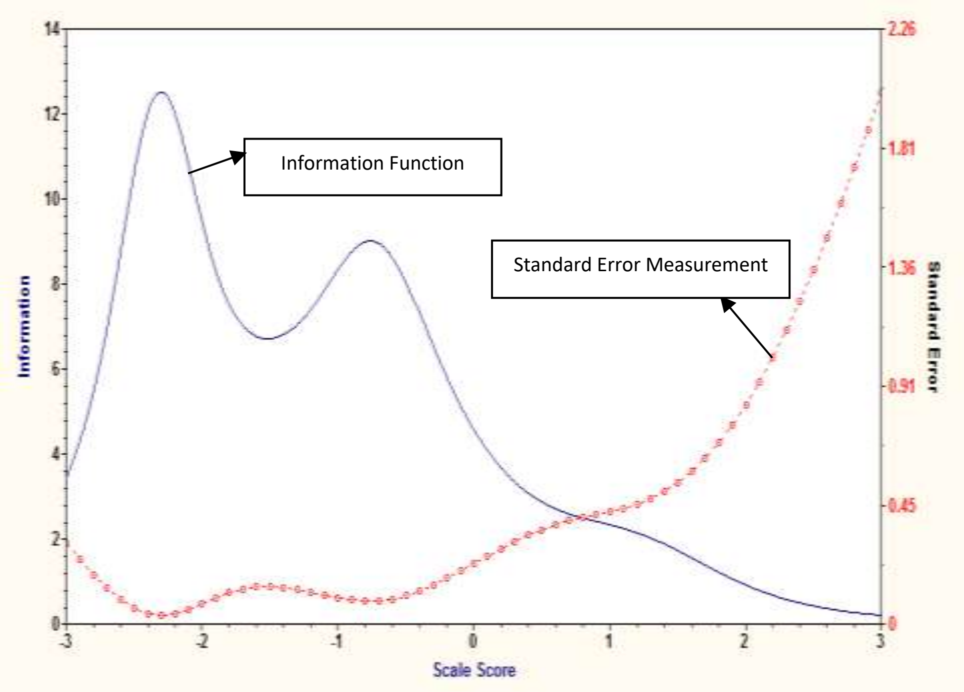

Figure 2. Information Function and Standard Error Measurement

Figure 2 showed that the test provided accurate information for students with ability estimate (theta) of -3.0 to 0.8. Even tests also provided accurate information for students with theta less than -3.0. The test provided the highest information for students with tetha around -2.4.

Discussion, Conclusion and Recommendations

Discussion

Instrument of assessment of knowledge aspect in the form of essay test produced by this study was proved empirically and theoretically to be in a good quality. Based on the validity of the contents and the validity of the construct, the entire test items were of good quality. The content validity index of V Aiken on the overall test item was 0.88-1.00 and the Porter alignment index was 0.93. This level of alignment belonged to very high category. This is because the test has been compiled based on Bloom's revised taxonomy as a reference to the applicable curriculum. The results of the alignment test strengthened the evidence of the validity of the content because it showed the suitability of the instrument items with the standard in the curriculum used.

The use of alignment index in this study will serve as a reference related to the validity of the contents of an assessment instrument. This study is an important one as revealed by Tindal (2005) who states that the results of alignment studies can be used to determine whether restructuring of the assessment is necessary or not. The use of alignment index on this study has provided more in-depth information about the quality of the test, which means that the test has alignment with the standards in the curriculum. It would be beneficial to use alignment index in other assessment instruments, especially in the form of tests, because information about alignment between assessment and standards is certainly very useful in policy making related to assessment and education. In addition, further development regarding alignment testing in assessment of mathematics activities can also be achieved in alignment testing between standards in curriculum and handbooks used by students. This is also suggested by Hasmi, Hussain, and Shoaib (2018) who reviewed the alignment between curriculum of mathematics and textbook using the SEC method.

All test items supported the test constructs. From the point of reliability coefficient, the test reliability also belonged to the good category. The test had good reliability with a CR coefficient of 0.88 and an Alpha coefficient of 0.78. It showed that the test had a high consistency of measurement results. However, one should pay attention to various factors affecting reliability in the implementation of this instrument for a large scale. As stated by Ebel and Frisbie (1991), if the test is used as a standard test, then, the reliability coefficient should be between 0.85-0.95 when used for the minimum class of 0.65. If test is used on a large scale and is intended for crucial decision, it should be standardized and has a reliability coefficient between 0.85-0.95. The tests produced in this study were more suitable for assessment within the scope of the class.

The quality of the test instrument in terms of its item characteristics was good. All test items were categorized as good with a discrimination index of 0.310-2.008 and item difficulty index of -2.329 to -0.475 in logit scale. From the points of the test information function and standard error measurement, the test provided accurate information on theta -3.0 to 0.8 and the highest information on the theta of about -2.4. It meant that the tests produced in this study were appropriate to be used for students with moderate and lesser abilities.

Conclusion and Recommendations

The essay test produced by this study was of good quality proven by several theoretical evidence supports and accurate empirical evidence. The result of this study is very useful, especially for Junior High School teachers because the result of the test can reveal achievement of students' statistical knowledge competence appropriately. In addition, for researchers and other educational practitioners, the result of this study is very important as it supports references regarding the quality of an assessment instrument. Further studies need to be conducted in other areas with larger and more diverse participants so as to further generalize and provide evidence to the quality of the instruments produced in this study. The alignment study also needs to be conducted with other methods such as Webb, SEC, or with other approaches such as online systems. Further studies can also be carried out by examining the use of the proposed item response of polytomous theory such as NRM, the response model for multiple-choice items, RSM, PCM, GRM, and sequential models for ordered responses in determining the characteristics of a test.

References

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational and Psychological Measurement*, 45, 131-142.

Ananda, S. (2003). *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd.

Anderson, D., Irvin, S., Alonzo, J., & Tindal, G. A. (2015). Gauging item alignment through online systems while controlling for rater effects. *Educational Measurement: Issues and Practice*, 34, 22-33.

Anderson, L. W. & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Bhola, D. S., Impara, J. C., & Buchendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice,* 22(3), 21–29.

Biggs, J. (2003). *Teaching for quality learning at university.* Glasgow: The Society for Research into Higher Education & Open University Press.

Buhaerah. (2010). Pengembangan perangkat pembelajaran berdasarkan masalah pada materi statistika di kelas IX SMP. Gamatika. Nomor 1. Nopember.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont: Wadsworth Group.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of test. *Psychometrika*, 16, 297-334.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement.* USA: Prentice-Hall Inc.

Effendi, K. N. S. & Farlina, E. (2017). Kemampuan berpikir kreatif siswa SMP kelas VII dalam penyelesaian masalah statistika. Jurnal Analisa. 3(2). 130-138.

Garson, G. D. (2009). *Overview structural equation modeling*, http://faculty.chass.ncsu.edu/garson/PA765/structur.htm.

Guler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90. http://dx.doi.org/10.14689/ejer.2014.55.5.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7thed). Prentice Hall [versi elektronik].

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory*. Newbury Park, CA: Sage Publication Inc.

Hanjarwati, R. & Wiyarno, Y. (2015). Pengembangan bahan ajar matematika (materi statistik) dengan menggunakan model active learning sistem 5 M untuk siswa kelas VII. *Jurnal Teknologi Pembelajaran Devosi.* 5(2).

Hasmi A., Hussain T., & Shoaib A. (2018). Alignment between Mathematics Curriculum and Textbook of Grade VIII in Punjab. *Bulletin of Education and Research,* April 2018, 40(1), 57-76.

Igbaria, M., Zinatelli, N., Cragg, P., & Cavaye, A. L. M. (1997). Personal computing acceptable factors in small firms: A structural equation model. *MIS Quarterly*, September, 279-299.

Kayapinar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research*, 57, 113-136, http://dx.doi.org/10.14689/ejer.2014.57.2.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement,* 17(4), 351-393.

Muraki, E. & Bock, D. (2002) PARSCALE 4.1 Computer program. Chicago: Scientific Software International, Inc.

Naga, D. S. (1992). Pengantar teori skor pada pengukuran pendidikan. Jakarta: Gunadarma

Nasstrom, G. & Henriksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Eletronic Journal of Research in Educational Psychology,* 6(3), 667-690.

Nitko, A. J. & Brookhart,S. M. (2011). *Educational assessment of students* (6th ed.). Boston, MA: Pearson Education Inc.

Nunnally, J. C. (1981). *Psychometric theory* (2nd ed). New Delhi: McGraw-Hill Publishing Company Limited.

Oriondo, L. L. & Antonio, D. E. M. (1998). *Evaluation educational outcomes.* Manila: Rex Printing Compagny.

Reeve, B. B., & Fayers, P. (2005). Appliying item response theory modeling for evaluating questionnaire item and scale properties. Dalam Fayers, P. & Hays, E.D. (Eds), *Assessing quality of life in clinical trials: Methods of practice* (2nd ed). New York: Oxford University Press.

Sireci, S. & Bond, M. F. (2014). Validity evidence based on test content. *Psicothema,* (26)1, 100-107.

Thorpe, G. L. & Favia A. (2012). *Data analysis using item response theory methodology: An introduction to selected programs and applications.* Psycology Faculty Scholarship. Paper 20. http://digitalcommons.library.umaine.edu/psy_facpub/20.

Tindal, G. (2005). *Alignment of alternate assessments using the webb system.* Washington, DC: Council of Chief State Officers.

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Walstad, W. B. (2006). Testing for depth of understanding in economics using essay questions. *Journal of Economic Education*. Washington: Winter.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education (Research monograph No. 6)*. Washington, DC: Council of Chief State School Officers.

Wells, C. S., Hambleton, R. K. & Purwono, U. (Juni 2008). Item response theory. Polytomous response IRT models and aplicatios. *Handout* delivered on the Training of Educational Assessment and Psychology (Psychometry), at Yogyakarta State University.

Wiggins, G. & McTighe, J. (2001). *Understanding by Design (2nd Ed.).* Alexandria, VA: Association for Supervision and Curriculum Development.

Wijanto, S. H. (2008). *Structural equation modeling dengan Lisrel 8.8*. Yogyakarta: Graha Ilmu.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum Associates, Inc.