ME – 53

THE COMPARISONS OF RELIABILITY ESTIMATION ON THE COMPOSITE SCORE OF MATHEMATICS TEST

Heri Retnawati
Yogyakarta State University, Indonesia
retnawati_heriuny@yahoo.co.id

Abtract

In a test, the final score was obtained as a composite of factors built up the test. The factors load in a test, would affect reliability of score of a test as composite of its factors. This study compared the reliability of the composite scores of mathematics test regardint its the loading factor. This study used the national examinations data of mathematics test in Junior High School, which was then analyzed using exploratory factor analysis to determine the loadingfactor in many cases basics on factors and long of the test. Furthermore the reliabilities estimated, and then compared with the H*-test and the Hakstian and Whalen test. The results showed that the analysis of 1-factor, the coefficient of reliability on a set of 20 items and a set of 25 items are higher than the original test load of 30 items. In the analysis of the two factors, the coefficient of reliability on a set of 15, 20, and 25 items are higher than the original test load of 30 grains when analyzed by one factor. In the analysis of the three factors, a set of 20 and 25 items have higher reliability coefficient compared with the original test load of 30 items when analyzed by one factor.

Key words: reliability estimation, composite score, mathematics test

In the development and management of test, reliability is one thing to be concerned. This reliabily known by looking at the coefficient of reliability of the score of test. The reliability coefficients can be interpreted as the coefficient of constancy or stability of the measurement results. A reliable instrument is capable to produce stable measurement results (Lawrence, 1994) and consistent (Mehrens & Lehmann, 1973: 102). The reliable instrument is said to have a high reliability coefficient when used to measure the same thing at different times the same or close to the same results. In this case, reliability is a nature of a set of scores (Frisbie, 2005). In relation to education, using a reliable instrument, the measurement results will be the same information though different raters, different tester or different items but measuring the same thing and have the same characteristics of the items.

Allen & Yen (1979: 62) stated that the test said to be reliable if the score of observations have a high correlation with the actual scores. Furthermore, it is stated that the reliability coefficient is the correlation coefficient between the two scores observations obtained from the measurement results using the parallel test. Thus, the definition of which can be obtained from the statement is a test that is reliable if the measurement results approach the actual state of the test participants.

In a study, it is usually used instruments involving many items. To understand this kind of data, typically used factor analysis. Factor analysis is used to reduce the data, to find relationships between independent variables (Stapleton, 1997), which is then collected in a smaller number of variables to determine the structure of the latent dimensions (Anonymous, 2001; Garson, 2006), which is called the factor. This factor is a new variable, which is also called the latent variable, the constructs variable and has properties can’t be observed directly (unobservable). In factor analysis, it is known the squared factor loading. The squared load factors stated magnitude of the variance in the observed variable that can be explained by factors (Van de Geer, 1971). The explained variance of the observed variables expressed as proportions, which is the ratio between the variance of these variables to the total variance of the overall observed variables. There are two types of factor analysis, the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA).

Ide dasar analisis faktor baik eksploratori maupun konfirmatori adalah mereduksi banyaknya variabel. Misalkan variabel awalnya adalah x_1, …, x_q, dan selanjutnya akan ditemukan himpunan faktor laten ξ_1, …, ξ_n (dengan n < q). Besarnya variabel yang dapat diamati (*observable*) merupakan hasil dari kombinasi linear faktor laten ξ_1 yang dinyatakan dengan

The basic idea of the both exploratory and confirmatory factor analysis are reducing the number of variables. Suppose the initial variables are x_1, …, x_q,, and will find a set of latent factors ξ_1, …, ξ_n (with n <q). The amount of variables that can be observed (observable) is the result of a linear combination of the latent factors ξ_1 expressed by

$$X_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + ... + \lambda_{in}\xi_n + \delta_i \text{ ....(1)}$$

In this case δ_i (measurement error) is a typical part of the x_i are assumed to be uncorrelated with $\xi_1, \xi_2, ..., \xi_n$. If $i \neq j$, $\delta_i \neq \delta_j$.

An Exploratory factor analysis is a technique for detecting and assessing latent source of variation or covarians in a measurement (Joreskog & Sorbom, 1993). An exploratory factor analysis is exploring the empirical data to locate and detect characteristics and relationships between variables without specifying the model to the data, in other words, look for the number of factors based on empirical data. In this analysis, the researchers do not have a priori theory to formulate hypotheses (Stapleton, 1997).

A confirmatory factor analysis is used to investigate the many factors that have been set previously supported by empirical data. This analysis is based on the premise that each manifest variable or observable variables can't completely describe a concept or a latent variable or construct variables. Related to this, on the basis of the theory, the concept of latent variable or variables or constructs can be described jointly by several manifest variables.

To determine the number of factors, by maintaining eigenvalues are more than one. These eigenvalues can be determined in several ways, the easiest is through the scree plot. The next tested whether the reduction factor or the addition of a significant factor of previous factor, using the difference between the value χ^2 obtained when placing k factor with when placing k + 1 factor (du Toit, 2003).

Pada skor komposit yang melibatkan faktor, untuk mengestimasi koefisien reliabilitas perlu didefinisikan model aditif teori tes klasik terlebih dahulu. Model aditif dinyatakan menjadi

In the composite score which involves factors, to estimate the reliability coefficient needs to be defined first additive model in classical test theory. Otherwise be additive model

$$\mathbf{X} = \mathbf{B}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \text{ ....(2)}$$

Where X is a vector of order n of observations scores, vektor $\boldsymbol{\tau}$ with order k of true scores, **B** is a matrix nxk that define the relationship between **X** and $\boldsymbol{\tau}$, $\boldsymbol{\varepsilon}$ is a measurement error vector. The measurement error and the true scores can not be obtained directly, but should be estimated. In accordance with the assumptions of the classical test theory, E ($\boldsymbol{\tau}$) = $\boldsymbol{\mu}$, E ($\boldsymbol{\varepsilon}$) = **0**, cov ($\boldsymbol{\tau},\boldsymbol{\varepsilon}$) = **0**,, and write var($\boldsymbol{\varepsilon}$) with $\boldsymbol{\Psi}$ ($\boldsymbol{\Psi}$ is a diagonal matrix), according Vehkalahti (2000: 21), the covariance structure of observation variable **X** is written by $\sum$ expressed by

$$\sum = \mathbf{BB'} + \boldsymbol{\Psi} \text{ ....(3)}$$

The reliability coefficient of factor scoores estimate using formula

ρ = **B' $\sum$ B** ....(4)

The results of the reliability coefficient estimates in equation (4) is a reliability coefficient factors measured in the test and still in matrix form.

To estimate the coefficient of reliability of the total score of X with k factor models, McDonald (1999), Kamata, Turhan, & Darandari (2003) and Bentler (2004) defines the coefficient of reliability as a proportion of the 'common' variance to the total variance. Furthermore, they prove the equation for the total score for estimating reliability coefficient of k-factor model in the equation:

$$\rho_{uu} = \frac{I_n^{'}\Lambda\Lambda' I_n}{I_n^{'}\Sigma I_n} \quad \text{...(5)}$$

ρ_{uu} is a reliability coefficient involving factors, Λ is a factor loading matrix, I_n is a vector with element 1 with order n, and Σ is the variance covariance matrix. Equation (5) is the reliability coefficient involving the factors that will be used to estimate the reliability coefficient in this study.

One way to increase the magnitude of the reliability coefficient is to extend the test, as long as the item is added to be homogeneous or measure the same thing. If the item is added is not homogeneous, the reliability coefficient of test item does not increase but instead, will decrease.

For the purposes of the election of the test, the test users to select tests that have higher reliability coefficient. To determine whether a test reliability coefficient is higher than the other test reliability coefficient, can be used to test the equality of two coefficients of reliability. The similarity of the two reliability coefficient can be determined by H*-test developed by Feldt (Feldt & Brennan, 1989). The value of H * to test the equality of two coefficients of reliability $\hat{r}_{x_1x_1'}$ and $\hat{r}_{x_2x_2'}$ and satisfies the equation

$$H^* = \frac{(1-\hat{r}_{x_1x_1'})}{(1-\hat{r}_{x_2x_2'})} \quad \text{...(6)}$$

With H* has F distribution with degrees of freedom $N_x - 1$ and $N_y - 1$. To compare two or more reliability coefficients, can be used to test Hakstian and Whalen m coefficients (Feldt & Brennan, 1989; Kim & Feldt, 2008). If n_ℓ the number of items and N_ℓ is the number of test takers, m is number of reliability coefficients were compared ($\ell = 1,2,3,...m$), and $\hat{r}_\ell$ the reliability to ℓ estimation results, the value for the Hakstian and Whalen test satisfies the equation 7.

$$M = \sum_{\ell=1}^{m} A_\ell - \frac{\left[\sum_\ell A_\ell (1-\hat{r}_\ell)^{-1/3}\right]^2}{\sum_\ell A_\ell (1-\hat{r}_\ell)^{-2/3}} \quad \text{...(7)}$$

where $A_\ell = \dfrac{(n_\ell - 1)(9N_\ell - 11)^2}{18(n_\ell)(N_\ell - 1)}$ and M in χ^2 distribution with m-1 degree of freedom.

By using response data of test takers, can be estimated reliability coefficient of set of items, consisting of 30, 25, 20 and 15 involving 1, 2 and 3 factors.

Method

This study uses a quantitative approach. The data document of student's responses to the mathematics test of national examinations consists of 30 items originally, which is then reduced based on the characteristics of the items and the results of focus group discussions. Data were analyzed using exploratory analysis to determine the factor loading. By using this factors, it is estimated the reliability and then compared with the H*- test and Hakstian and Whalen test.

Results

The results of the factor analysis of the adequacy of the sample shows the value of Chi-square test is 21863.839 Bartlet with 435 degrees of freedom and p-value less than 0.01. These results indicate that the sample size of 3,012 is used in this study is in adequate category.

Table 1
KMO and Bartlett Test Result

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.962
Bartlett's Test of Sphericity	Approx. Chi-Square	21863.839
	df	435
	Sig.	.000

Based on the results of the factor analysis using SAS/IML, it can be obtained that the students' mathematics response data to the national examination had 4 eigen values greater than 1, so it can be said that the test load 4 factors. Of these four factors, there are 59.14% of variance that can be explained. Furthermore, the significance of these factors was tested by using χ^2 test.

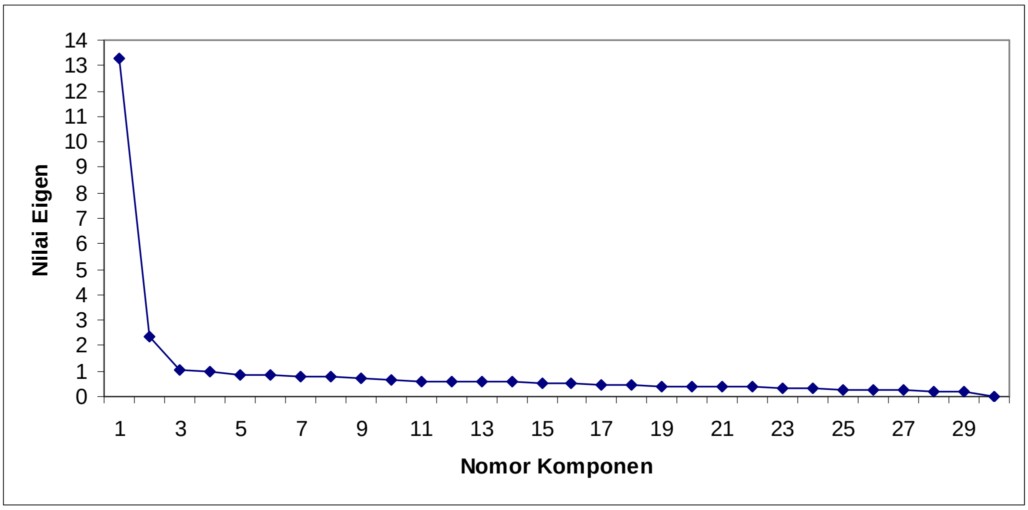


Figure 1
Scree Plot of Exploratory Factor Analysis

Eigen values can be presented with the scree plot in Figure 1. Looking at the results of the scree plot, eigen values are ranging ramps appear on the 3rd factor. It shows that there is one dominant factor on the math test, 1 other factors also contributed substantially to the components of variance that can be explained. Starting the third factor, and so on, the graph shows already began to plateau. This indicates that the device measures the math test at least 2 factors with the first factor is the dominant factor.

Another way that can be done to determine the number of factors is contained by comparing the chi-squared value of each factor on factor analysis. The χ^2 value in this analysis is computed with the help of the TESTFACT program. By conducting factor analysis by including only 1 factor, the value of Chi-square and degrees of freedom sebesesar 33353.97

2951.00 for. On the second factor, the value of Chi-square and degrees of freedom of 33124.35 2922.00 and the third factor of the degrees of freedom 2894.00 33006.25. Lastly, entering the 4 factors in this analysis will be obtained χ^2 value of the degrees of freedom 2867.00 36387.73. Furthermore, the difference can be calculated chi-squared values to determine which model is better. More test results are presented in Table 2.

Based on the results of this analysis, it can be concluded that based on empirical data, the mathematics test in national examinationss device with a better measure consecutive three factors, two factors, and one factor. These factors, here in after referred to as dimensions. The results of this statistical test based on the outcome of the determination of the factors by using the scree-plot, which shows that the tests measure two-dimension, but for the purposes of this study, there are three dimensions that become variables, that are 1,2, and 3 dimensions.

Table 2
Result of χ^2-test to Determine Loading Factor

k Factor	χ^2	df	$\chi^2_{(k)}$- $\chi^2_{(k+1)}$	$df_{(k)}$- $df_{(k+1)}$	$\chi^2_{kritis\ (0,05,\ df)}$	Conclusion
1	33353,97	2951				
2	33124,35	2922	229,62	29	42,56	2 factor model is better than 1 factor model
3	33006,25	2894	118,10	28	41,34	3 factor model is better than 2 factor model
4	36387,73	2867	-3381,48	27	40,11	4 factor model isn’t better than 1 factor model

Based on the results of determining the number of factors is contained, then it is performed the naming of factors. Naming the factors were done based the loading factor after rotated, with regard to the loading factor of more than 0.4. Naming factors contained in the test conducted by researcher with the help of mathematicians, practitioners (2 teachers), mathematics education expert and psychologist in forum Focus Group Discussion (FGD). Previous analysis with 2 factors and analysis by incorporating three factors use promax rotation. This rotation includes the rotation nonortogonal category. This is done because in model 2 factors, first and second correlation factor is 0.3559, while the 3-factor model, the first and second correlation factor is 0.3110, the first and the third factor is 0.3457, and the second with the third factor is 0.3069. Furthermore, experts named the factor based on the loading factor of each item that is more than 0.4.

Based on the results of the FGD, in the 1 factor model, the factor was named the general math skills. For model 2 factors, the first factor was named general ability and factors and the second was named spatial ability, while for the three factors, the first factor was named general ability, spatial ability was name of the second factor and the third factor was named with a numerical ability. The names of these factors are based on the loading factors as the results of the factor analysis after rotation nonortogonal.

The first factor in model 2 factors named by the general mathematical ability caused by the loading factor rotation results in the first factor includes the overall minimum basic competencies that should be achieved by test takers. The second factor was named spatial abilities because 4 of 5 items that have a load factor of more than 0.4 were items associated with the spatial ability that were angle in triangle, triangle comparison, circle, trigonometric and other items about logarithms.

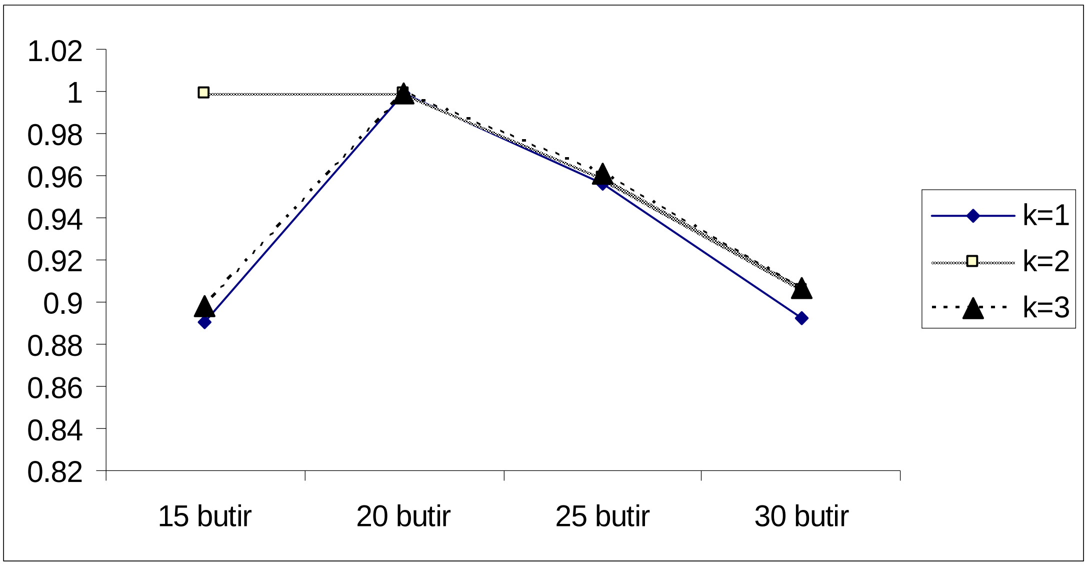
On models with 3 factors, the first factor were named by the general math skills. The second factor is named with spatial abilities caused by the loading factor which is more than 0.4

are items about the angle of the triangle, the ratio of triangle, circle, Pythagoras and the area of a triangle, trigonometry, and logarithms. On the third factor, called numerical ability caused by the loading factor which is more than 0.4 contained in items that also require numerical ability to solve, namely the set of numbers, the angle of the triangle, the nature of parallel lines, area of parallelogram, straight line equation, and the comparison triangle.

Once the factors named, then the reliability of scor are estimated. The estimation results are presented in Table 3. To understand the patern, the result is shown in Figure 2.

Table 3
The Result of Reliability Coefficient Estimation

Model	30 butir	25 butir	20 butir	15 butir
1 dimesion	0,8920	0,9565	0,9991	0,8901
2 dimesion	0,9054	0,9589	0,9992	0,9995
3 dimesion	0,9070	0,9611	0,9992	0,8985



K=1 (1 dimension model), K=2 (2 dimension model), K=3 (3 dimension model)

Figure 2
The Result of Reliability Coefficient Estimation

Looking at the graph, it appears that in the 3-factor model, the reliability coefficient was slightly higher than in model 1 factor. In model 2 factors, the item pool consisting of 15 and 20 items, reliability coefficient is higher when compared to models 1 and 2 factors, but the item collection consisting of 25 and 30 items, the results of a reliability coefficient estimate were similar to the model 3 factor.

To test the similarity of 12 reliability coefficient is used Hakstian and Whalen test. The hypothesis (H_0) tested is the reliability of twelve coefficients are equal. From the calculation, the value of M is 64196.3995 and test = 19.68. These results indicate that H_0 is rejected, which means the proficiency level of the twelfth reliability coefficient, whose value is not the same. Further test H * to test whether the coefficient of reliability of the test with a reduced items or tests are analyzed with models 2 and 3 factors better than the reliability coefficient of the initial test load of 30 items and analyzed with model 1 factor. The results of the analysis are presented in Table 4.

Looking at the results in Table 4, it can be obtained that the analysis of the first factor, the coefficient of reliability on a set of 20 items and a set of 25 items are higher than the original test load of 30 items. In the analysis of the two factors, the coefficient of reliability on a set of

15, 20, and 25 items are higher than the original test load of 30 items when analyzed by one factor. In the analysis of the three factors, a set of 20 and 25 items have higher reliability coefficient compared with the original test load of 30 items when analyzed by one factor.

Discussion

Observing the results of exploratory factor analysis in scree-plot, it is found that the mathematics test of national examination not only measure the dominant factor, but also measure other factors. If the only measured one dominant factor, the main contribution of this factor is only about 44.29% of the total variance explained. This contribution is still relatively far from the figure of 100%, a number that is expected by the test developers, in explaining the variation in the ability of test takers. The main contribution of this factor can be improved by adding other factors that inluded measured in the mathematics.

Observing eigen values as result of factor analysis, found that there are four eigenvalues are more than 1. This indicates, there may be four factors that can contribute a large proportion of the total variance that can be explained. However, these four factors are not necessarily significant when included as a factor that is contained in the mathematics test. By Chi-square test, can be obtained that the factor analysis model which contains two factors better than the load factor analysis model 1 factor. Similarly, 3-factor model, which is a better model than the model contains two factors, but the model includes four factors are not better than the models with 3-factor analysis. Based on these results, we can conclude that there are three factors contained in the mathematics test of national examination in 2006.

Table 4

The Equaity Test of Reliability Coefficients

Cases	ν_1; ν_2	H* (to 30 items 1 faktor)	F_{tabel}, α=5%	Conclusion (Equaty of Reliability)	**Interpretasi** Compared with reliability coeffisients of 30 items 1 faktors
15 items 1 faktor	30;15	0.982712	2.25	Not Rejected	Equal
20 items 1 faktor	30;20	120	2.04	Rejected	Higher
25 items 1 faktor	30;25	2.482759	1.87	Rejected	Higher
30 items 1 faktor					
15 items 2 faktors	30;15	216	2.25	Rejected	Higher
20 items 2 faktor	30;20	135	2.04	Rejected	Higher
25 items 2 faktors	30;25	2.627737	1.87	Rejected	Higher
30 items 2 faktors	30;30	1.141649	1.84	Not Rejected	Equal
15 items 3 faktors	30;15	1.064039	2.25	Not Rejected	Equal
20 items 3 faktors	30;20	135	2.04	Rejected	Higher
25 items 3 faktors	30;25	2.77635	1.87	Rejected	Higher
30 items 3 faktors	30;30	1.16129	1.84	Not Rejected	Equal

By inserting one more factor in the preliminary analysis models into analysis with 2 factors, there was an increase in total variance that can be explained. The second contribution of this factor in explaining the variance of 52.19%, in other words there is a contribution of 7.90% rise compared to only entering one factor in the analysis. If the models plus one more factor that into 3 factors, the measured variance contributions be 55.76% or increasing in contribution of 3.57%. Noting these results, the largest variance contributed by the first factor alone, while the contribution of the second and third factors in explaining the variance is smaller than the first factor.

In factor analysis, the first eigenvalues is a greatest value compared with another eigenvalues. Magnitude eigenvalues shows a linear dependence on the data. On the second factor, the third and so on, eigenvalues quite small compared to the first eigenvalues (Johnson & Wicern, 2002). Because of the large variance that can be explained by a factor proportional to the magnitude of eigenvalues, then the first factor in the analysis of the factors contributing to the greatest compared to other factors.

Discuss more about the factors, there are three factors that measured at the mathematics test in 2006. This means that the test as mathematics measure at least 3 factors of ability, which in this study is defined as a dimension. In accordance with the load factor after rotated, these factors later named. In accordance with the recommended materials experts, the first factor, a factor which is named with the general mathematical ability, the two factors named by the general ability and spatial and on 3 factors, factors named by the general ability, spatial, and numerical. These results indicate, there is another dimension that measured in the mathematics test, or in other words the mathematics test can measure the ability of the one-dimensional general ability, the ability of two-dimensional-general and spatial ability, and 3 dimensional abilities that are general ability, spatial, and numerical.

The results of the analysis in this study showed the mathematics test measured more than one dimension or contain multidimensional. These results are supported by a statement Reckase (1997), Bolt & Lall (2003), Ackerman, Gierl, & Walker (2003) and strengthen the results of research studies conducted by Thulber, Shinn, & Smolkowski (2002), that the learning achievement test measures more than one dimension. Similarly Badrun Kartowagiran & Heri retnawati (2007) which showed that the national examination mathematics test in 2003 and 2005 measured more than one dimensions.

Based on the similarity coefficient of reliability test results, it can be obtained that the analysis of the first factor, the coefficient of reliability on a set of 20 items and a set of 25 points higher than the original test load of 30 items. In the analysis of the two factors, the coefficient of reliability on a set of 15, 20, and 25 items higher than the original test load of 30 items when analyzed by one factor. In the analysis of the three factors, a set of 20 and 25 items are higher reliability coefficient compared with the original test load of 30 items when analyzed by one factor. This can be explained that the analysis at 20 and 25 in model 1 factor, 2 factors, and 3 factors, can be obtained reliability coefficient better than the original test reliability coefficient. That are understandable because it reduces the items from 30 to 25, then from 25 to 20, in addition to considering the content also consider the quality of the item, or discard the items are not well in advance. With good grain, which can be explained variance will be larger than when estimating the reliability coefficient original test containing both good items and not good items.

Conclusions

The results showed that the analysis of 1-factor, the coefficient of reliability on a set of 20 items and a set of 25 items are higher than the original test load of 30 items. In the analysis of the two factors, the coefficient of reliability on a set of 15, 20, and 25 items are higher than the original test load of 30 grains when analyzed by one factor. In the analysis of the three factors, a set of 20 and 25 items have higher reliability coefficient compared with the original test load of 30 items when analyzed by one factor.

Referencies

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Anonim. (2001). Factor analysis. *Journal of Consumer Psychology*, Vol. 10(1&2), 75-82.

Badrun Kartowagiran & Heri Retnawati. (2007). Validasi dimensi perangkat tes ujian nasional 2003-2006. *Laporan Penelitian*. Lembaga Penelitian UNY.

Croker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehard and Winston Inc.

Du Toit, M. (2003). *IRT from SSi: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood: SSi.

Ebel, R.L. & Frisbie, D.A. (1986). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Feldt, L.S. & Brennan, R.L. (1989). Reliability. Dalam *Educational measurement* (3thed). (Linn, R.L. (Ed.)). London: Collier Macmillan Publishers.

Garson, D. (2006). *Factor analysis*. Diambil dari http://www2.chass.ncsu.edu/garson/pa765/index.htm tanggal 24 September 2006.

Joreskog, K. & Sorbom, D. (1996). *Lisrel 8*. Chicago: Scientific Software International Inc.

Joreskog, K. & Sorbom, D. (1993). *Lisrel 88: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Scientific Software International.

Kamata, A., Turhan, A., & Darandari, E. (2003). Estimating reliability for multidimensional composite scale scores. *Paper presented at annual meeting of American Educational Research Association, Chicago., April 2003.* Diambil dari http://www.coe.fsu.edu aera Kamata2.pdf pada tanggal 1 Oktober 2007.

Kim, S. & Feldt, L.S. (2008). A comparison of tests for equality of two or more independent alpha coefficients. *Journal of Education and Measurement*, Vol. 45. No. 2. pp. 179-193.

Lawrence, M.R. (1994). Question to ask when evaluating test. *Eric Digest*. *Artikel*. Diambil dari: http://www. ericfacility. net/ ericdigest/ ed.385607.html tanggal 6 Januari 2007.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Elrbaum.

Mehrens, W.A. & Lehmann, I.J. (1973). *Measurement and evaluation in education and psychology*. New York: Hold, Rinehart and Wiston,Inc.

Nunally, J. (1978). *Psychometric theory (2nd ed.)* . New York: McGraw Hill.

Stapleton. (1997). *Basic concepts and procedures of confirmatory factor analysis*. Diambil dari http://ericae.net/ft/Cfa.HTM tanggal 25 September 2006.

Van de Geer, J.P. (1971). *Introduction to multivariate analysis for the social sciences*. San Francisco, VA: W.H. Freeman and Company.