Estimation of Cronbach reliability based on sample size, gender, and the grades

Ikhsanudin[1,2], Bambang Subali[1], Heri Retnawati[1], Edi Istiyono[1]
[1]Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2]Department of Mechanical Engineering Education, Faculty of Teacher Training and Education, Universitas Sultan Ageng Tirtayasa, Serang, Indonesia

Article Info

Article history:

Received Jun 30, 2022
Revised Jan 29, 2023
Accepted Feb 9, 2023

Keywords:

Creativity test
Cronbach reliability
Gender
Grade of class
Sample size

ABSTRACT

The study tries to see the effect of different sample sizes on the reliability of creativity tests using Cronbach's alpha. In addition, the background of the sample (gender and grade) is also detected. The study has a practical significance for designing a sample for estimating reliability using Cronbach alpha. The secondary data are used for analysis. The sample of creativity test is determined purposively, that is, has been administered to 1,960 students. The α-coefficient calculation uses the SPSS program. The result proves in the total sample, the creativity test is reliable (α=0.852). In small sample sizes, the α-coefficients values vary greatly in the different test groups. In the larger sample size, the α-coefficients are closer to constant in the different test groups. For a polytomous scale test with a length of 20 items, the α-coefficients obtained close to constant on the sample size of more than 80 testees. This result indicates for the optimum result the sample size is at least four times larger than the number of test items in reliability analysis, especially Cronbach alpha. Based on the gender aspect, descriptively the α-coefficient in the male group (0.846) was almost equal to the female group (0.851). By grouping the testees based on the grades, there are slight differences in α-coefficients in class IV (0.850), class V (0.841), and class VI (0.809). Regarding the results of this study, in the reliability estimation of test development, it is suggested to consider the adequacy and the characteristic of samples.

Corresponding Author:

Ikhsanudin
Department of Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta
Caturtunggal, Depok, Sleman, Special Region of Yogyakarta 55281, Indonesia
Email: ikhsanudin.2019@student.uny.ac.id/ikhsanudin@untirta.ac.id

1. INTRODUCTION

The categorization of educational purposes is the central attention of curriculum developers. The purpose of education in the narrow sense leads to learning objectives. Anderson and Krathwohl emphasized that the goals of education and learning are important to discuss [1]. In addition to being the basis of curriculum development, learning objectives are also the basis for the development of assessment instruments. Subali explains that the assessment serves to collect data or information to determine the level of achievement of learning objectives that have been designed and implemented [2]. In this case, the objectives of learning can be divided into three dimensions: cognitive, affective, and psychomotor which are formulated in the form of standard competencies. This competency is assessed by using a prescribed procedure.

The achievement of students' competence on the cognitive dimension is generally assessed by the test. Azwar described the test as a question set devised to uncover certain attributes through response to the question [3]. Mardapi added that the test is one form of an instrument used to measure competency [4]. The questions composed of a test have the category of right/wrong answers according to the competence or objectives to be achieved. Formally, one of the learning objectives in the cognitive dimension is to increase the creativity of a child [5]. The ability to think creatively is a psychological attribute that is latent (hidden) and cannot be recorded directly. Therefore, measurement experts in the field of education develop creativity tests to obtain information about the creativity of children ranging from early childhood to college level.

Creativity tests are developed by experts in a variety of forms, including short-field tests and essay tests. Both of the tests are also called supply response testing and are often used to measure high-order thinking skills [2]. In this case, creativity is a manifestation of the high-order thinking ability of students that needs to be developed especially through the learning process. The results of creativity tests provide data or information about the level of children's creative thinking ability. To be able to carry out the creativity test, measurement instruments are needed based on the existing indicators. To achieve a high degree of objectivity toward children's creativity, the measurement instruments require an accurate and trusted procedure. The results of creativity measurements can be interpreted appropriately if the data obtained from good-quality instruments. One of the criteria of a good instrument is reliability. In test development, reliability is required to ensure that the test is of good quality and capable of delivering reliable results [5], [6].

The results of creativity test development have been widely published in both print and electronic media. Each of the creativity test instruments has advantages and disadvantages. Criticism of the test is required for the refinement of the creativity test at a later time [7]. Although creativity tests are commonly used, they still need to be reviewed on an ongoing basis to investigate their quality [8]. One of the things that need to be criticized in the measurement of creativity is reliability [9]. Reliable test instruments for measuring child creativity can still be re-explored in subsequent reliability [10]. The reliability of creativity tests needs to be analyzed and proven on an ongoing basis [11].

Reliability on creativity tests can be determined based on a classical test theory approach by applying one of several formulas that have been developed. The classical test theory uses a simple mathematical model in estimating the reliability of a test based on an analysis of the relationship between the observational score, a true score, and an error score. However, this classical test theory approach contains weaknesses in that the results of the measurement depend on the characteristics of the test, and the grain parameters depend on the ability of test participants. In addition, measurement errors in classical test theories can only be searched for groups of testees [4]. In other words, different groups of testees allow different results of test analysis. Reliability is quite sensitive to the bias of test results [12]. In this case, the heterogeneity of the testee groups allows the bias of the test results.

Concerning the different groups of the testee, each child taking the test is a distinct individual, so the members of the test group are almost certainly varied, for example in terms of gender and grade. The gender aspect is a common thing that is often studied in psychological issues where male and female have their stereotypes [13]. In late childhood, the cognitive aspect of female is superior to male [14]. Furthermore, the grade of class shows the different psychological development, especially in the cognitive dimension of the child where the higher grade ideally has a higher cognitive ability. In this case, the analysis of test results vertically should be given attention to reliability [15].

The reliability estimate is highly dependent on the characteristics of the sample following the tests [16]. Based on the foregoing description, there is a high probability that there is a difference in test reliability estimation results in terms of gender and grade level. In addition, since the classic reliability determination depends on the characteristics of the group of testees, the effect of the number of testees against the reliability of the creativity test is also necessary to investigate. The sample size can affect the standard size error of the correlation coefficient included in the reliability analysis [17]. The results of Lee and Zhang's study concluded that one of the factors affecting the reliability estimate is the sample size [18]. Therefore, the sample size (the number of testees) that is taking the creativity test is also likely to influence the reliability estimation results in the test.

Creativity as one of the learning goals is very important to investigate its development from an early age using a developed test instrument. Given the importance of creativity and test reliability as a criterion of measurement instruments and giving attention to the number of the testee, gender, and grade level of class as a variation of characteristics of the group of the testee, the estimation of reliability on creativity tests based on sample size, gender, and grade level are important to investigate. The reliability estimation in this research is expected to complete the understanding of the reliability and become the basis for the teacher and or researcher related to the development of test instrument and reliability analysis as well as the meaning of the result, especially on the creativity test and other test using the polytomous scale. This study aims to estimate the reliability of creativity measurement based on gender and grade of class as well as taking concern the sample size of students taking the test. The calculation of the reliability coefficient uses the Cronbach formula.

2. RESEARCH METHOD

This study focuses on the results of a creativity test of a science process skill which consists of 20 open-ended items and has been administered to 1,960 elementary school students in the Special Region of Yogyakarta, Indonesia. The sample of the test involved participants in which 647 students of grade IV, 660 students of grade V, and 653 students of grade VI from 84 schools in over five areas in 14 sub-area, with six schools samples in each area. The sample participants were determined by purposive sampling technique; determined based on certain considerations. Grade IV, V, and VI students are selected as samples because they are assumed to be students who have experience in science learning that enhance the creativity at the elementary school and represent the learning achievement of the school.

The data of the creativity test results in this study is secondary data derived from Subali's research on the creativity of the science process skills of elementary school students. The test instrument used has been validated and standardized and it is one of four sets of creativity tests so the testee analyzed amounted to about a quarter of the total students. The data about gender and grade of students were obtained using questionnaires filled out by students on test answer sheets. Administration of the test is carried out in cooperation with supervisors and teachers. Student data and measure results through the test are recorded according to the relation scores of individual test items and individual students. This framework provides the basic data for the analysis of the reliability of the measurements investigated. The effect of sample size, gender, and grade of class against the reliability is analyzed by grouping the samples according to the number of testees, gender, and student's grade and then analyzing the reliability in each group of testees.

Analysis of creativity test results uses the scoring guidelines with a polytomous scale (scores of 0 to 6) for each item by the rubric. If the testee successfully answers all the test questions correctly, then the maximum score is gained for 120. After each item is scored by all students, then it is made in the relation framework of items scores and samples of the testee and analyzed the reliability of scores using the Cronbach alpha formula used SPSS for Windows. The analysis of sample size effects on the creativity test reliability estimation was carried out in a multilevel organized sample group starting from the school level, the cluster of schools (three schools), sub-city level (six schools), city level (area), and provincial level. Thereafter, groupings of students were sampled along with the scores of each test item by gender and grade to estimate reliability for each group of male and female students, as well as the student grade IV, V, and VI. The dependability analysis of gender and grade of class to reliability is also done in stages. The results of the analysis are the values of α-coefficient which are presented descriptively in cross-tables.

3. RESULTS AND DISCUSSION

3.1. Information of the test and the initial reliability of measurement used for the reliability studies

The creativity test referred to in the title of this study is a form of creativity measurement in life science process skills that have been tested on elementary school students. The instrument of the test has been validated empirically using the Rasch model which utilizes the quest program. This research uses some of the data of test results to examine more deeply the reliability estimates of the results of measurements in terms of sample size, gender, and grade level. The main results show the reliability estimate of the scores from creativity measurement is the α-coefficient which is obtained from computation using the Cronbach formula. Broadly speaking, the α-coefficient on the measurement is presented in Table 1.

From Table 1, it can be seen that the α-coefficient on the creativity test scores on the total sample studied is in the high category. This proves that the test is reliable. In addition, the value of the α-coefficient in each group of testees is not below 0.8. Thus, it can be said that the reliability of this creativity test is good so the scores obtained from the creativity test can be trusted to the ability of students.

Table 1. Cronbach α-coefficient on the creativity test for groups of testees in overall

Gender	Grade						Total	
	IV		V		VI			
	α	N	α	N	α	N	α	N
Male	0.842	323	0.830	332	0.800	345	0.846	1000
Female	0.849	324	0.841	328	0.810	308	0.851	960
Total	0.850	647	0.841	660	0.809	653	0.851	1960

α=Cronbach's alpha coefficient; N=number of testees

3.2. The Cronbach reliability based on sample size, gender, and grade of class

Taking into the small number of student samples at the class level, the division of students based on the gender in Table 2 is not conducted. Based on Table 2, it is recognized that the value of α-coefficients is very diverse. The range of α-coefficients is also high at 0.787 (0.128-0.915), at the school level. In the school cluster level, the range of the α-coefficient becomes smaller that is equal to 0.239 (0.648-0.887). This suggests that the reliability estimation using the Cronbach formula for the creativity test with these groups have not satisfying yet, so sample expansion is required to improve the accuracy of the estimation.

In Table 3, the range of α-coefficient for sub-city and city levels is lower than in the school-level estimate and is a little different from the school clustering level which is presented in Table 2. At the sub-city level, by considering the gender and grade, the α-coefficients range from 0.667 to 0.926 (range of 0.259). The lowest reliability is on the creativity test scores for the group of female students of grade VI and the highest in the scores of the fourth-grade female student group. Based on the gender, the results of creativity test reliability estimation at the sub-city and city level are not much difference between male and female students group, the greatest difference is equal to 0.100. Furthermore, based on the grade, it can be seen that there is a greater difference in the reliability estimation result which is 0.259. This result indicates that the α-coefficient on the creativity test is more influenced by class level than gender. On the other side, estimating the reliability using the α-coefficient is more effective to control the gender than the grade. To obtain a more satisfying results, it is necessary to extend the sample regarding the gender and grade of the class.

Table 2. Cronbach α-coefficient for groups of testee in schools cluster of school level

School name	Grade						Total	
	IV		V		VI			
	α	N	α	N	α	N	α	N
ES A	0.152	7	0.792	6	0.689	7	0.785	20
ES B	0.771	4	0.716	6	0.128	7	0.791	17
ES C	0.757	7	0.537	5	0.161	5	0.523	17
ES Cluster 1	0.648	18	0.887	17	0.753	19	0.884	54
ES D	0.712	7	0.711	7	0.641	7	0.889	21
ES E	0.567	4	0.641	4	0.558	4	0.893	12
ES F	0.248	4	0.915	3	0.641	7	0.766	14
ES Cluster 2	0.684	15	0.741	14	0.697	18	0.858	47
Sub-area 1	0.913	33	0.838	31	0.680	37	0.872	101

α=Cronbach's alpha coefficient; N=number of testees

Table 3. Cronbach α-coefficient for groups of testee in sub-city and city level

Location	Gender	Grade						Total	
		IV		V		VI			
		α	N	A	N	α	N	α	N
Sub-area 1	Male	0.880	16	0.813	15	0.669	16	0.857	47
	Female	0.926	17	0.713	16	0.667	21	0.862	54
Sub-area 2	Male	0.880	48	0.809	45	0.812	62	0.845	155
	Female	0.852	57	0.845	55	0.792	44	0.849	156
Area 1	Male	0.888	64	0.828	60	0.789	78	0.855	202
	Female	0.875	74	0.827	71	0.757	65	0.851	210
	Total	0.881	138	0.842	131	0.774	143	0.855	412

α=Cronbach's alpha coefficient; N=number of testees

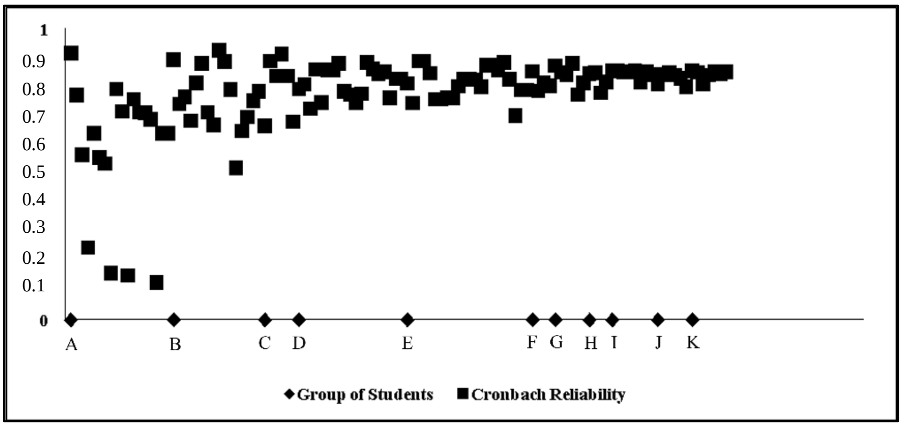
Increasing the sample size gives a positive impact on the reliability estimation of children's creativity measurement results using Cronbach α-coefficient. Generally, Table 4 presents the α-coefficient of the larger sample size in the smaller range of difference than the level below. Based on gender, the reliability at the provincial level is said to be almost constant with the highest difference of α-coefficient only 0.011. Based on the grade of class, the difference in reliability is more pronounced which is the highest difference of 0.041. This demonstrates the higher effect of grade levels than gender on the reliability of creativity tests. However, overall, on large sample sizes, the values of α-coefficients in the groups are high (≥0.800).

Furthermore, based on the all tables (Tables 1 to 4), we can obtain 116 values of α-coefficients. Considering that students in the field naturally cannot be tightly controlled, the analysis can be focused only on the sample size so that it can be made the distribution of α-coefficient based on sample size. From Figure 1, it can see that on the small sample size, the α-coefficient are highly different in ranges. As the sample size is enlarged, the difference in ranges is smaller until it approaches the constant values in the large sample size (>80). The distribution of the α-coefficient based on the sample size demonstrates that estimating the reliability of a test requires sufficient samples to avoid biased coefficients either under-estimate or over-estimate results. The adequacy of sample size for the estimated reliability of Cronbach α-coefficient in this case for creativity tests or general tests may be studied further, in particular, considering the homogeneity or heterogeneity of the testees.

Table 4. Cronbach α-coefficient for groups of testee in province level

City/Area	Gender	Grade						Total	
		IV		V		VI			
		α	N	α	N	α	N	α	N
Area 1	Male	0.888	64	0.828	60	0.789	78	0.855	202
	Female	0.875	74	0.827	71	0.757	65	0.851	210
Area 2	Male	0.813	89	0.804	93	0.800	73	0.842	255
	Female	0.802	67	0.702	77	0.827	76	0.817	220
Area 3	Male	0.858	74	0.854	83	0.786	84	0.854	241
	Female	0.884	76	0.825	71	0.762	66	0.855	213
Area 4	Male	0.746	46	0.776	51	0.744	62	0.780	159
	Female	0.763	59	0.847	64	0.764	65	0.817	188
Area 5	Male	0.744	50	0.725	45	0.773	48	0.813	143
	Female	0.785	48	0.860	45	0.837	36	0.850	129
Province	Male	0.842	323	0.830	332	0.800	345	0.846	1000
	Female	0.849	324	0.841	328	0.810	308	0.851	960
	Total	0.850	647	0.841	660	0.809	653	0.852	1960

α=Cronbach's alpha coefficient; N=number of testees



Description: Notations A through K show the sample size groups

A≤10
B=11-20
C>20
D>40
E>60
F>80
G>100
H>150
I>200
J>300
K>400

Figure 1. The diversity of α-coefficients based on sample size

3.3. Discussion

The reliability of test scores, as estimated through measures of internal consistency, has been characterized mathematically in many ways that appear, on the surface at least, to be very dissimilar to one another [19]. From the literature review, it can be obtained information that the reliability of Cronbach alpha (α-coefficient) is one of the most common forms of internal consistency in the measurement of the social field. The α-coefficient can provide a guarantee or empirical proof that the test result (scores) of a measurement process can be trusted so that the score can give the right meaning to the student's ability, in this research is the creativity of elementary school students. Traub and Rowley say that the scores on a test can be used as measurements of the variable or to make decisions [20]. In this case, on the basis that the results of creativity tests carried out are reliable, then some criteria interpretation of the score as the ability of students are fulfilled.

The accuracy of the meaning of the scores depends on the accuracy (validity) of the test. Test validity is supported by reliability on the same test. Reliability is a precursor to test validity [21]. Reliability contributes to ensuring the effectiveness of measurement [22]. Thus, reliability needs to be estimated and reported to support the validity of the measurement, in this case including the measurement of the children's creativity through the test. Based on the reliability estimates presented in Table 1, it is evident that the results of creativity tests are generally reliable with the α-coefficient of 0.851. The internal consistency reliability according to Wells and Wollack for the lower-stakes standardized tests is greater than or the same as 0.80 [21]. The coefficient on the creativity test in this study has met the category. As a form of reliability estimation, the internal consistency of a measure can be a focal point in an evaluation of the potential adequacy of an instrument for adaptation to another cultural setting. Cronbach's alpha (α-coefficient) is often used as the statistical index for such a determination. However, alpha presumes a τ-equivalent of the test and may constitute an inaccurate population estimate for multidimensional tests [23]. Cronbach's alpha is analytically handicapped to capture important measurement errors and scale dimensionality, and how it is not invariant under variations of the scale length, inter-item correlation, and sample characteristics [24].

In some cases, any formula to estimate reliability cannot be said to be perfect, including the α-coefficient also has some drawbacks. First, the α-coefficient always has a value that cannot be equal to the test score's reliability given the inter-item covariance matrix and the usual assumptions about measurement error. Second, in practice, the α-coefficient is used more often as a measure of the test's internal consistency than as an estimate of reliability. However, it is shown clearly that the α-coefficient is unrelated to the internal structure of the test. It is necessary to be further discussed that statistics based on a single test administration do not convey much information about the accuracy of individuals' test performance [25]. The assumptions underlying the α-coefficient are unlikely to hold in practice, and violation of these assumptions can result in nontrivial negative or positive bias. Structural equation modeling was discussed as an informative process both to assess the assumptions underlying the α-coefficient and to estimate the reliability [26]. When the assumptions are met, the average value of the alpha coefficient is very close to the population reliability level. Despite having some weaknesses, the α-coefficient is still accepted and used as a reliability estimator by experts. The results of Wigley's study reveal that three popular myths about Cronbach's alpha are dispelled because: i) it does not measure the dimensionality of a test; ii) it is not a test characteristic; and iii) it is not always the best estimate of reliability [27]. In other words, the existence of Cronbach α-coefficient is still very good today as one of the ways to get the proof of reliability. However, there is still a need for a study of several factors that may affect the α-coefficient in reliability.

Factors that are likely to affect reliability estimation results with α-coefficients, especially on creativity tests attempted to be revealed in this study include sample size, gender, and grade of the class. Theoretically from the literature review, these three factors can give indirectly influence reliability estimation based on the effect on the variation of the ability scores (creativity scores) in the groups of testees. Tables 2, 3, 4, and Figure 1 provide an overview of the α-coefficient variation in the values estimated based on sample size, gender, and grade level. Concerning the sample size at a certain grade, the number of students who take the test is less than ten children, and at the school level around 20 children so in this small sample, the reliability estimate provides a high-different range of α-coefficient. At this level, according to the sample adequacy criterion, it can be said that the reliability estimation using the Cronbach alpha formula cannot be done. Therefore, it is necessary to enlarge the sample size.

Tables 2 to 4 show that enlarging the sample has a positive effect on the α-coefficient of the creativity test, the larger the sample size the smaller the different range of α-coefficient. Based on the results presented in Figure 1, the small different ranges begin in a sample group with a size of more than 80 testees. Thus, it takes more than 80 respondents to estimate the reliability of the creativity test with a polytomous scale in 20 items length. The determination of sample size adequacy is an important factor in reliability analysis [28]. The results can be studied in further research by considering the heterogeneity characteristics of the testee in specific means, especially for the analysis of sample adequacy. Related to Cronbach's reliability, Feldt and Kim stated that the α-coefficient is a special type of correlation [29]. In this case, the sample size has effects on correlation results [30]. Thus, this statement is in line with the results of this study, it can be said that the α-coefficient is influenced by the sample size, particularly in the small sample size.

Furthermore, the gender and grade of class effects were analyzed, that reliability needed to be considered in test scores analysis both horizontally and vertically [15]. In this case, the gender is horizontally analyzed, and the grades are vertically. According to the gender, during the child's cognitive development, there is a difference in ability between male and female cognitive aspects [13], [14], [30]. This gender (gender stereotype) allows the test result bias. The results of the analysis in Tables 3 and 4 showed that the α-coefficient values are almost equal between the male and female students. This indicates that controlling the gender positively impacts the estimated reliability that is likely to occur due to a decrease in the bias of the test results because the reliability is quite sensitive to the test bias [12].

Based on the study result, it is known that there is dependability between sample size, gender, and grade level toward the estimation of Cronbach reliability on creativity tests for elementary school students. The study of factors affecting the estimation of reliability, including those studied in this study, is still possible to be studied in further research by focusing on specific variables and controlling the others. This research does not rule out the possibility that the reliability of measurement especially on creativity tests can also be influenced by the other variables. The factors that may cause the biases of the test result are also likely to affect the reliability. The testee's response toward the test to estimate reliability is difficult to avoid bias. Other factors against the reliability that can be studied in further research besides this study are the test length, the scale category used, the differential item functioning (DIF) items, item types, levels of group ability, and response categories [18]. These factors likely affect the value of the test reliability. Furthermore, a study of these variables in reliability testing in general or especially in creativity tests can be performed either separately or simultaneously. The results of the study are expected to provide a complete and comprehensive description of the reliability estimates, particularly using the Cronbach alpha formula.

4. CONCLUSION

The result of the creativity test with a polytomous scale that has been administered to elementary school students is proved reliable (α=0.852). Based on the sample size, the reliability α-coefficients began to stabilize at the cluster of school or sub-area level with the total number of testees being more than eighty students. Thus, in estimating the reliability of the creativity test in twenty-item length scoring polytomous, it is encouraged to use samples of over eighty respondents. In general, the larger the sample size the smaller the difference ranges of α-coefficient, or in other words the α-coefficients are more stable in the larger group of the testee, for the optimum result the sample size at least four times larger than the number of test items in reliability analysis, especially Cronbach alpha. The estimation of Cronbach's reliability should be considered the sample adequacy. In terms of the gender and grade level, the α-coefficient on the male test group (0.846) was relatively similar to that of the female test group (0.851) and there was a slight difference of α-coefficients in the groups of students grade IV group (0.850), grade V (0.841), and grade VI (0.809).

The results of this study indicated that there is a difference in reliability in creativity tests for elementary school students viewed by gender and grade level where grade level has more influence than gender. Furthermore, a more in-depth study of the acceptable limits of sample size for reliability estimation especially on creativity tests concerning the homogeneity of the characteristics of the testee so that the samples for estimation are more constant and still meet the minimum sample criteria. The study can also be done based on the number of test items and the difference of the polytomous scale used in the test and other factors. It is possible, these are the weaknesses of the Cronbach reliability either in creativity testing or in reliability estimation in general.

REFERENCES

[1] L. Anderson and D. R. Krathwohl, *The foundation framework for learning, teaching, and assessment*. Yogyakarta, Indonesia: Pustaka Pelajar (in Indonesian), 2017.

[2] B. Subali, *Development of tests and investigations of validity and reliability empirically*. Yogyakarta, Indonesia: UNY Press (in Indonesian), 2016.

[3] S. Azwar, *Constructing psychological scale*. Yogyakarta: Pustaka Pelajar (in Indonesian), 2017.

[4] D. Mardapi, *Measurement, assessment, and evaluation of education*, Yogyakarta: Parama Publishing (in Indonesian), 2017.

[5] Law Number 20 of 2003 concerning the National Education System. Jakarta, Central Government of the Republic of Indonesia (in Indonesian), 2003.

[6] S. Azwar, *Reliability and validity*. Yogyakarta: Pustaka Pelajar (in Indonesian), 2012.

[7] K. H. Kim, "Is creativity unidimensional or multidimensional? Analyses of the Torrance tests of creative thinking," *Creativity Research Journal*, vol. 18, no. 3, pp. 251–259, Jul. 2006, doi: 10.1207/s15326934crj1803_2.

[8] K. H. Kim, "Can we trust creativity tests? A review of the Torrance tests of creative thinking (TTCT)," *Creativity Research Journal*, vol. 18, no. 1, pp. 3–14, Jan. 2006, doi: 10.1207/s15326934crj1801_2.

[9] D. Hocevar, "Measurement of creativity: review and critique," *Journal of Personality Assessment*, vol. 45, no. 5, pp. 450–464, Oct. 1981, doi: 10.1207/s15327752jpa4505_1.

[10] N. Stefanic and C. Randles, "Examining the reliability of scores from the consensual assessment technique in the measurement of individual and small group creativity," *Music Education Research*, vol. 17, no. 3, pp. 278–295, Jul. 2015, doi: 10.1080/14613808.2014.909398.

[11] E. Rudowicz, D. Lok, and J. Kitto, "Use of the Torrance tests of creative thinking in an exploratory study of creativity in Hong Kong primary school children: a cross-cultural comparison," *International Journal of Psychology*, vol. 30, no. 4, pp. 417–430, Aug. 1995, doi: 10.1080/00207599508246577.

[12] E. Peer and E. Gamliel, "Too reliable to be true? Response bias as a potential source of inflation in paper-and-pencil questionnaire reliability," *Practical Assessment, Research, and Evaluation*, vol. 16, no. 1, 2011.

[13] J. W. Santrock, *Educational psychology*, Jakarta: Salemba Humanika (in Indonesian), 2014.

[14] J. W. Santrock, *Child development period, Book 2,* 11th Ed. Jakarta, Indonesia: Salemba Humanika (in Indonesian), 2011.

[15] A. Allalouf, G. Klapfer, and M. Fronton, "Comparing vertical and horizontal scoring of open-ended questionnaires," *Practical Assessment, Research, and Evaluation*, vol. 13, no. 1, 2019.

[16] B. Sumintono and W. Widhiarso, *Rasch modeling application on educational assessment*, Cimahi, Indonesia: Trim Komunikata (in Indonesian), 2015.

[17] A. Pfaffel and C. Spiel, "Accuracy of range restriction correction with multiple imputation in small and moderate samples: A simulation study," *Practical Assessment, Research, and Evaluation*, vol. 21, no. 1, 2016.

[18] Y.-H. Lee and J. Zhang, "Effects of differential item functioning on examinees' test performance and reliability of test," *International Journal of Testing*, vol. 17, no. 1, pp. 23–54, Jan. 2017, doi: 10.1080/15305058.2016.1224888.

[19] H. Li and H. Wainer, "Toward a coherent view of reliability in test theory," *ETS Research Report Series*, no. 1, Jun. 1998, doi: 10.1002/j.2333-8504.1998.tb01751.x.

[20] R. E. Traub and G. L. Rowley, "Reliability of test scores and decisions," *Applied Psychological Measurement*, vol. 4, no. 4, pp. 517–545, Oct. 1980, doi: 10.1177/014662168000400406.

[21] C. S. Wells and J. A. Wollack, "An instructor's guide to understanding test reliability." Testing and Evaluation Services University of Wisconsin, 2003.

[22] D. Bacon, "The contributions of reliability and pretests to effective assessment," *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, 2004.

[23] S. Thurber and Y. Kishi, "Coefficient alpha and interculture test selection," *Assessment*, vol. 21, no. 2, pp. 250–252, Apr. 2014, doi: 10.1177/1073191112444921.

[24] M. A. Bujang, E. D. Omar, and N. A. Baharum, "A review on sample size determination for Cronbach's alpha test: a simple guide for researchers," *Malaysian Journal of Medical Sciences*, vol. 25, no. 6, pp. 85–99, 2018, doi: 10.21315/mjms2018.25.6.9.

[25] M. Amirrudin, K. Nasution, and S. Supahar, “Effect of variability on Cronbach alpha reliability in research practice,” *Jurnal Matematika, Statistika dan Komputasi*, vol. 17, no. 2, pp. 223–230, Dec. 2020, doi: 10.20956/jmsk.v17i2.11655.

[26] S. B. Green and Y. Yang, “Commentary on coefficient alpha: a cautionary tale,” *Psychometrika*, vol. 74, no. 1, pp. 121–135, Mar. 2009, doi: 10.1007/s11336-008-9098-4.

[27] C. J. Wigley, “Cronbach’s alpha versus components of variance approach (COVA): dispelling three myths about alpha and suggesting an alternative reliability statistic for communication trait research,” *Communication Research Reports*, vol. 28, no. 3, pp. 281–286, Jul. 2011, doi: 10.1080/08824096.2011.591220.

[28] I. Kennedy, “Sample size determination in test-retest and Cronbach alpha reliability estimates,” *British Journal of Contemporary Education*, vol. 2, no. 1, pp. 17–29, Feb. 2022, doi: 10.52589/BJCE-FY266HK9.

[29] L. S. Feldt and S. Kim, “Testing the difference between two alpha coefficients with small samples of subjects and raters,” *Educational and Psychological Measurement*, vol. 66, no. 4, pp. 589–600, Aug. 2006, doi: 10.1177/0013164405282488.

[30] I. Trizano-Hermosilla and J. M. Alvarado, “Best alternatives to Cronbach’s alpha reliability in realistic conditions: congeneric and asymmetrical measurements,” *Frontiers in Psychology*, vol. 7, May 2016, doi: 10.3389/fpsyg.2016.00769.

BIOGRAPHIES OF AUTHORS

Ikhsanudin is a doctoral student in Educational Research and Evaluation in Universitas Negeri Yogyakarta, Indonesia. He is also a lecturer in Mechanical Engineering Education, Faculty of Teacher Training and Education, Universitas Sultan Ageng Tirtayasa, Indonesia. His research focuses on the development of educational measurement instrument, educational assessment, and teacher education. He can be contacted at email: ikhsanudin.2019@student.uny.ac.id or ikhsanudin@untirta.ac.id.

Bambang Subali is a Professor in Biology Education in Universitas Negeri Yogyakarta, Indonesia. His area of expertise is educational assessment, research methodology, and statistics. His research interest is related to measurement instrument development, learning innovation, and biology education assessment. He can be contacted at email: bambangsubali@uny.ac.id.

Heri Retnawati is a Professor in Educational Research and Evaluation in Universitas Negeri Yogyakarta, Indonesia. The focus of her research interests is related to mathematics learning innovation, mathematics education assessment, and psychometrics. She can be contacted at email: heri_retnawati@uny.ac.id.

Edi Istiyono is a Professor in Educational Research and Evaluation in Universitas Negeri Yogyakarta, Indonesia. He is a head of the doctoral program of Educational Research and Evaluation. His expertise in measurement, assessment, and evaluation in physics education. The focus of his research interests is related to higher order thinking skills of students and assessment in physics education. He can be contacted at email: edi_istiyono@uny.ac.id or edi_istiyono_uny@yahoo.co.id.