Available online at: http://journal.uny.ac.id/index.php/pythagoras
PYTHAGORAS: Jurnal Matematika dan Pendidikan Matematika, 18 (2), 2023, 97-111

Trends of Stability of Reliability Coefficient Based on Sample Size and Ability of Test-takers

Busnawir [1*], Kodirun [2], Zamsir [3], Hafiluddin Samparaja [4], Hasnawati [5]
Department of Mathematics Education, Universitas Halu Oleo, Kendari, Indonesia
* Corresponding Author. E-mail: busnawir@uho.ac.id

ARTICLE INFO

Article History:
Received: 17-Mar. 2023
Revised: 07-Mar. 2024
Accepted: 18-Mar. 2024

Keywords:
Ability of the test takers, sample size, stability of the reliability coefficient

ABSTRACT

One aspect that needs to be considered in the assessment of learning outcomes is the quality of the test by a stable reliability coefficient. This study aims to determine the trend of the stability of the reliability coefficient of the mathematics formative test based on the sample size and the ability of the test takers. The study was experimental in the form of a simulation, using a population of scores based on the answers of 403 test takers. The research sample was taken from the population of scores with 19 variations in sample sizes. Each sample size was repeated 31 times with the return technique; the reliability coefficient was calculated for each repetition and was used as the unit of analysis. In addition to the differences in sample sizes, the differences in the abilities of the test takers were also seen in two categories of high and low. Data were analyzed using exploratory-descriptive statistics and analysis of variance. Results showed as follows: first, the formative test of mathematics that was developed by the teacher at school has a reliability coefficient in the inadequate category; second, the reliability coefficient of the test tends to be more stable with increasing sample sizes; third, the difference in the ability of the test takers does not make a significant difference to the reliability coefficient; fourth, there is no interaction between sample sizes and abilities of the test takers on the reliability coefficient of the test.

How to Cite:
Busnawir, B., Kodirun, K., Zamsir, Z., Hafiluddin, S., & Hasnawati, H. (2023). Trends in the stability of the test reliability coefficient: an overview based on sample size and ability of test-takers. *Pythagoras: Jurnal Matematika dan Pendidikan Matematika, 18*(2), 97-111. https://doi.org/10.21831/pythagoras.v18i2.59392

https://doi.org/10.21831/pythagoras.v18i2.59392

INTRODUCTION

Assessment of learning outcomes is carried out to measure the level of competency achievement of students (Shavelson et al., 2018; Zlatkin-Troitschanskaia et al., 2018), as a report on the progress of learning outcomes to improve the learning process further (Kuh & Ewell, 2010; Mustopa et al., 2021). The results of the assessment can later be used as a reference for teachers to map students' abilities (Brown & Harris, 2014; Hamilton et al., 2021) as well as an evaluation material for teachers (Suchyadi et al., 2020; Primasari et al., 2021; Retnawati et al., 2016). In terms of assessing the progress of student learning outcomes, consistent, systematic, and programmed measurements are required (Phillips & Phillips, 2016), by using good quality tests (Adom et al., 2020; Yuniartik et al., 2017).

In general, in schools, there are two kinds of tests that are often used to measure test takers' learning outcomes, namely formative and summative tests (Jian & Shaoqian, 2014; Ariawan et al., 2022), however, until now, there are still many learning outcome instruments used by teachers that do not meet the requirements of proper and good quality tests (Arum et al., 2022; Herman et al., 2021); there are still many complaints that arise in schools related to assessment (Idrus, 2022; Kartowagiran & Jaedun, 2016; Oktadini et al., 2022), a number of teachers still have a low ability in preparing good learning outcome tests (Kasli et al., 2022; Nopriyeni et al., 2019), Therefore, the quality of learning outcome tests developed by teachers in schools still needs to be studied (Ndiung & Jediut, 2020; Singh & Sarkar, 2015). There are still many teachers who do not understand in compiling item grids

(Ariawan et al., 2022), and analysis of the quality of learning outcome tests (Makhrus, 2018) even though the quality of the test is one aspect that needs to be considered in evaluating learning outcomes (Suardipa & Primayana, 2020; Jusrianto et al., 2022). A good quality test will provide more accurate information about the competency of the test taker (Muluki, 2020; Umami et al., 2021) and increase motivation and learning achievement (Iskandar & Rizal, 2018; Mudanta et al., 2020).

Learning outcomes tests should accurately measure students' abilities in order to minimize measurement errors, have a good level of accuracy (Antara et al., 2020; Arif, 2016; Dewi et al., 2019), and have the smallest possible error (Sinaga, 2016). The characteristics of an appropriate test instrument are related to validity and reliability (Junika et al., 2020). Reliability is an important factor in determining whether a test is good or not (Heale & Twycross, 2015), and expresses measurement results that are precise and can be convincing (Guna et al., 2014; Arum et al., 2022). Reliability indicates the degree of consistency of an instrument that underlies measurement errors that may occur in a testing process (Hidayat et al., 2017; Komperda et al., 2018).

Reliability relates to the stability of measurement results (Gunartha, 2022; Jalilibal et al., 2021; Khumaedi, 2012). Test reliability can be affected by scoring techniques and the number of alternative answers (Ariyanti & Bhakti, 2020; Bhakti, 2015; Reiter-Palmon et al., 2019), number of test items, group variability, estimation method, group level (Chalmers et al., 2016), and, including the ability level of the test takers (Hikamudin & Hairun, 2021). It is also necessary to pay attention to the group level in the test instrument trial because it can affect the reliability of the test (Magdalena et al., 2020). To increase the reliability of a test, it is necessary to create questions that are able to distinguish between test takers who are less intelligent and those who are proficient (Sarwanto et al., 2020; Nuriyah, 2014; Osborne et al., 2016). Apart from the ability factor of the test takers, another problem that often arises related to the quality of an instrument is determining the sample size of the instrument trial (Alwi, 2015). The sample size is an element of a research design that researchers need to consider in planning studies (Burmeister & Aitken, 2012). However, large sample sizes that must be used usually become an obstacle for educational practitioners in calibrating the developed instruments (Kummerfeld & Rix, 2019).

So far, there is no definite measure of the sample to determine the stability of the reliability coefficient of a test instrument. Several previous studies suggest that reasonable precision for reliability estimates requires approximately 50 study participants and at least 3 trials (Hopkins, 2000). The study conducted by Magdalena et al., (2020) used a sample size of 100 test takers in developing 30 items of anxiety instruments and obtained a high reliability coefficient of around 0.80. The development of an instrument using 10 times the number of items has been described by Nunnally (1970, 1986) as suggested by Afif et al., (2021); Atamimi (2014). On the other hand, Sapnas dan Zeller (Ono, 2020) concluded that a sample size of 50 is sufficient to evaluate psychometric properties on social construct measures. Argianti & Retnawati (2020) found that with a sample of 239 test takers' answers to 30 school exam questions, the reliability coefficient was obtained in a fairly high category. The results of research conducted by Taherdoost (2016) state that the nature of the data determines the sample size: the stronger the data the smaller the sample size required to obtain accurate results. Kennedy (2022); Suciati et al. (2020) states that for the stability and reliability of an instrument, a minimum of 200 respondents are required.

In the current study, experiments were carried out using a simulation approach based on empirical data. What is new and different from previous studies is that this study uses 19 variations in sample size, ranging from a sample size of 36 (3 times the number of items) up to a sample size of 252 (21 times the number of items). For each sample size, 31 repetitions were carried out; differentiate the abilities of the test takers in the high and low categories; and assess the stability of the reliability coefficient based on sample size and ability of test takers. The aims of this study were: (1) to determine the quality of the formative tests in mathematics which are developed by the teacher based on the test reliability coefficient criteria; (2) to know the correlation between sample size and reliability coefficient to find a representative sample size for the magnitude of the reliability coefficient; (3) to find a representative sample size for the stability of the reliability coefficient of the test relative to the number of items; (4) and to find out whether or not the difference in the level of ability of the test takers affects the magnitude of the reliability coefficient and also the interaction between the sample size and the ability of the test takers.

METHOD

The study was designed in the form of an experiment using a simulation method based on empirical data. Simulations were used to draw research samples repeatedly using the Minitab for Windows computer application program (Arisandi & Dewi, 2016; Cassettari et al., 2012; Eck & Liu, 2008). The research population consisted of the scores of the math formative exam results which were the answers of 403 test takers at State Junior High School 9 Kendari for the 2022/2023 academic year. The research score population was then divided into two categories: namely high ability and low ability based on the scores of the test results. The total scores of the test takers were sorted from lowest to highest, then divided into two groups; the lowest 50% were declared as low ability and the highest 50% high ability. Based on empirical data, test takers with low ability had a score range of 2-7, while high ability 8-12.

The samples for the study were taken from the population in varying sizes (nineteen sub-samples). Sample sizes, ranged from 3 times the number of items to 19 times the number of items, produced sample sizes of 36 (S-36), 48 (S-48), 60 (S-60), 72 (S-72), 84 (S -84), 96 (S-96), 108 (S-108), 120 (S-120), 132 (S-132), 144 (S-144), 156 (S-156), 168 (S- 168), 180 (S-180), 192 (S-192), 204 (S-204), 216 (S-216), 228 (S-228), 240 (S-240), 252 (S-252), hereinafter referred to as sub-samples 1, 2, 3, ..., 19. Sampling was done randomly using a statistical application (Minitab for Windows); for each sub-sample, it was repeated by way of replacement (Gillenwater et al., 2019; Taherdoost, 2016). Repetitions were carried out 31 times, referring to the normal distribution (Rapono et al., 2019). Furthermore, each repetition was calculated for the reliability coefficient, so that a total of 31 reliability coefficients were obtained for each sub-sample, and this was the unit of analysis. The design of the analysis unit is shown in Table 1.

Table 1. Unit of analysis design

Test Taker's Ability	Repetition to	Sample Size						
		1	2	3	...	17	18	19
		S-36	S-48	S-60	...	S-228	S-240	S-252
High (H)	1	$r_{1.1T}$	$r_{1.2T}$	$r_{1.3T}$	...	$r_{1.17T}$	$r_{1.18T}$	$r_{1.19T}$
	2	$r_{2.1T}$	$r_{2.2T}$	$r_{2.3T}$	...	$r_{2.17T}$	$r_{2.18T}$	$r_{2.19T}$
	...	...	...	...	...	...	...	...
	31	$r_{31.1T}$	$r_{31.2R}$	$r_{31.3R}$	...	$r_{31.17R}$	$r_{31.18R}$	$r_{31.19R}$
Low (L)	1	$r_{1.1R}$	$r_{1.2R}$	$r_{1.3R}$	...	$r_{1.17R}$	$r_{1.18R}$	$r_{1.19R}$
	2	$r_{2.1R}$	$r_{2.2R}$	$r_{2.3R}$	...	$r_{2.17R}$	$r_{2.18R}$	$r_{2.19R}$
	...	...	...	...	...	...	...	...
	31	$r_{31.\,1R}$	$r_{31.\,2R}$	$r_{31.\,3R}$	...	$r_{31.\,17R}$	$r_{31.\,18R}$	$r_{31.\,19R}$

Description :
r1.1T = reliability coefficient in the 1st repetition, 1st sample, high ability category.
r31.19T = reliability coefficient in the 31st repetition, 19th sample, high ability category
r1.1R = reliability coefficient in the 1st repetition, 1st sample, low ability category.
r31.19R = reliability coefficient in the 31st repetition, 19th sample, low ability category

The instrument used in this study was an objective test of learning outcomes of mathematics consisting of 12 items. The test was developed by the teacher and was used to measure the results of formative tests for the odd semester of 2022/2023. The test was held online by the school, data on the test results were collected via a Google form. The tests were carried out in shifts in the school laboratory under the teacher's supervision.

The reliability coefficient of the test result score was calculated using the KR-20 (Bajpai & Bajpai, 2014; Utami et al., 2022). The reliability coefficient was calculated based on the different sample sizes and the abilities of the test takers in the high-low category (Livingston, 2018). The resulting reliability coefficient data were then analysed using descriptive-explorative statistics and the analysis of variance. The descriptive-explorative statistics were used to see the trend of the reliability coefficient and the coefficient of variation produced by each sample size. The analysis of variance was used to determine whether there was a difference in the reliability coefficient between the sample sizes and the abilities of the test takers and the interaction between the two, at a significance level of $\alpha = 0.05$ (Kumar & Misra, 2020; Sianturi, 2022).

To determine the stability of the reliability coefficient, the magnitude of the coefficient of variation was used (Canchola, 2017), with the criterion that a large coefficient of variation describes large fluctuations, while a small coefficient of variation describes small fluctuations. For certain purposes, it is desired that small fluctuations are indicated by a small coefficient of variation. (Jalilibal et al., 2021; Calif & Soubdhan, 2016). A small coefficient of variation indicates a smaller error rate (Hadinata, 2018), and also describes data that is increasingly homogeneous (Hadinata, 2018; Kapantow et al., 2017). Homogeneous data reflects relatively the same characteristics as a characteristic of stability (Dewi et al., 2019; Schiel et al., 2018).

RESULTS

Based on the descriptive-explorative and analysis of variance analyses, the results are described and explained in this section. The results of the exploratory-descriptive analysis describe the statistical characteristics coefficient variation (CV) of the reliability coefficient values for 19 different sample sizes as summarized in Table 2.

Table 2. Confidence interval and the coefficient of variation of the reliability coefficient based on the sample size

No.	Sample Size	N	Mean	SE Mean	95% Confidence Intervals	Coefficient of Variation
1	S-36	31	0.599	0.014	(0.570797; 0.626622)	12.71
2	S-48	31	0.622	0.014	(0.593964; 0.649455)	12.17
3	S-60	31	0.596	0.009	(0.576327; 0.614835)	8.81
4	S-72	31	0.616	0.008	(0.599985; 0.631241)	6.92
5	S-84	31	0.608	0.009	(0.589169; 0.625863)	8.23
6	S-96	31	0.655	0.008	(0.638696; 0.670724)	6.67
7	S-108	31	0.645	0.009	(0.626333; 0.663602)	7.88
8	S-120	31	0.654	0.008	(0.637251; 0.671007)	7.03
9	S-132	31	0.639	0.008	(0.622759; 0.656080)	7.10
10	S-144	31	0.651	0.007	(0,636907; 0.665286)	5.94
11	S-156	31	0.642	0.007	(0.627049; 0.656628)	6.28
12	S-168	31	0.658	0.006	(0.645433; 0.670116)	5.12
13	S-180	31	0.657	0.006	(0.644844; 0.669220)	5.06
14	S-192	31	0.642	0.006	(0.628801; 0.654489)	5.46
15	S-204	31	0.610	0.004	(0.602510; 0.618070)	3.48
16	S-216	31	0.594	0.006	(0.582009; 0.606443)	5.61
17	S-228	31	0.601	0,005	(0.589782; 0.612153)	5.07
18	S-240	31	0.608	0.005	(0.598677; 0.617581)	4.24
19	S-256	31	0.597	0.004	(0.589292; 0.604902)	3.56

Table 2 shows that there is a tendency for the coefficient of variation of the reliability coefficient to decrease as the sample size increases. The smallest coefficient of variation was achieved at a sample size of 204 (17 times the number of test items), meaning that the larger the sample size, the more stable the reliability coefficient. The tendency of the average reliability coefficient and the coefficient of variation according to the size of the sample is increasingly clear in Figure 1 and Figure 2.

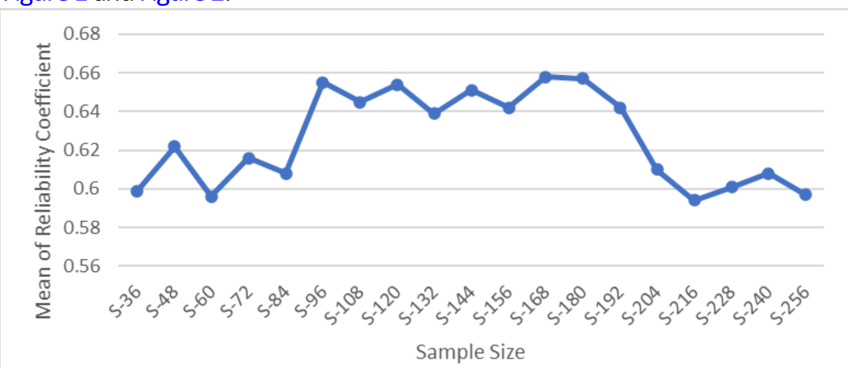

Figure 1. Average reliability coefficient based on the sample size

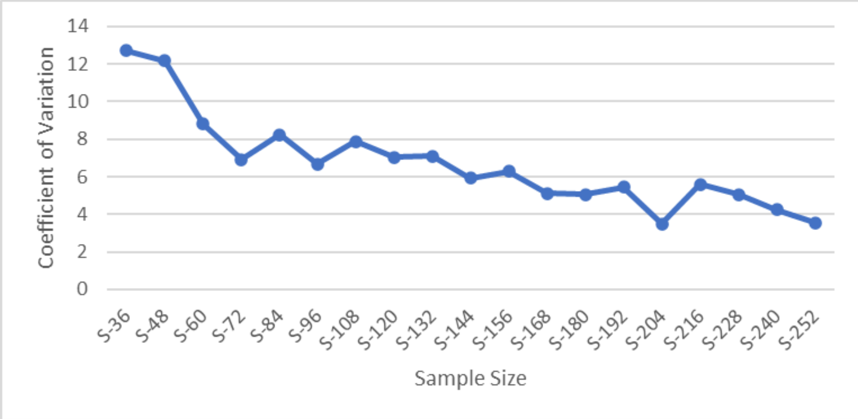

Figure 2. Coefficient of variation based on the sample size

In Figure 2 it can be seen that the magnitude of the coefficient of variation is in the range of 3.48 – 12.71, the smallest coefficient of variation is produced by a sample size of 204 (S-204) or 17 times the number of items, while the largest coefficient of variation is produced by a sample size of 36 (S-36) or 3 times the number of items. The magnitude of the coefficient of variation tends to decrease or get smaller when the sample size is enlarged.

Graphically, the coefficient of variation of the reliability coefficient values based on the sample size and the ability of the test takers (high and low) is shown in Figure 3. Visually, Figure 3 shows that the magnitude of the coefficient of variation produced by test takers with high and low ability fluctuates relatively the same and tends to get smaller as the sample size increases. However, at a very small sample size (S-36), it can be seen that the coefficient of variation for the abilities of the two participants appear to be different. This provides a strong indication that the coefficient of variation of the test reliability coefficient is not affected by differences in the test taker group. Other information that can be obtained is that a very small coefficient of variation can be achieved at a sample size of 204 (17 times the number of items).

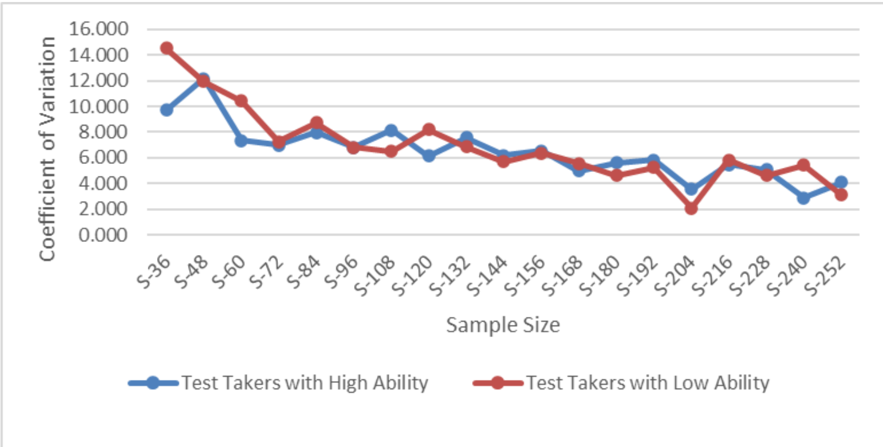

Figure 3. Coefficient of variation based on sample size and ability of test takers

The results of the test of the average difference in the reliability coefficient among the sample sizes of the study showed no significant difference as can be seen in Table 3. The test results summarized in Table 4 mean that there are differences in the reliability coefficient due to differences in sample sizes, which is indicated by the *p*-value = 0.000 <0.05 (5% confidence level). 22.63% of the resulting variance in the reliability coefficient was due to differences in sample sizes. The results of the different tests on the average reliability coefficient seen from the difference in sample size and the ability of the test takers are summarized in Table 4.

Table 3. Difference test results of average coefficient of reliability based on sample size

Source	*DF*	SS	MS	*F*	*p*
Sample Size	18	0.317	0.018	8.950	0.000
Error	551	1.084	0.002		
Total	569	1.401			
S = 0.044	R-Sq = 22.63%				

Table 4. Average difference test results for reliability coefficients based on sample size and AoTT

Source	*DF*	SS	MS	*F*	*p*
Students Ability (AoTT)	1	0.004	0.004	1.950	0.180
Sample Size	18	0.317	0.018	9.260	0.000
Students Ability*Sample Size	18	0.034	0.002	0.970	0.496
Error	532	1.046	0.002		
Total	569	1.401			

Table 4 explains that the difference in AoTT does not make a significant difference to the reliability coefficient, which is indicated by the *p*-value = 0.180 (greater than the 5% confidence level). The difference in sample size produces a significant difference in the reliability coefficient indicated by the *p*-value = 0.000 (< 5% confidence level). There is no interaction effect between sample size and AoTT on the reliability coefficient as indicated by the *p*-value = 0.496 (> 5% confidence level). There is 25.33% of the resulting variance in the reliability coefficient that was caused by differences in sample size and AoTT.

The test results using the *t*-test of two independent samples yield a *t*-Value = -0.42 and *p*-Value = 0.679 > *Alpha* = 0.05, so it is not significant at the 95% confidence level. This test indicates that there is no significant difference in the coefficient of variation of the reliability coefficient between high AoTT and low AoTT. However, quantitatively, the average value of the coefficient of variation for high AoTT is relatively smaller than that of low AoTT, each of which is 6.477 for high AoTT and 6.830 for low AoTT. Visually, the boxplot of the coefficient of variation of the two AoTT is shown in Figure 4.

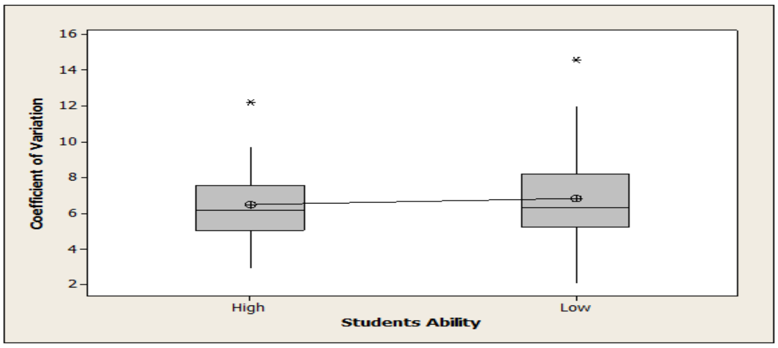

Figure 4. Boxplot of high and low AoTT variation coefficients

DISCUSSION

The results of the descriptive analysis provide an overview of the characteristics of the test reliability coefficient based on differences in the sample size and AoTT. The average test reliability coefficient did not tend to increase at sample sizes of 36, 48, 60, 72, and 84; experienced an increase in sample sizes of 96, 108, 120, 132, 144, 156, 168, 180, 192; then decreased again with sample sizes of 204, 216, 228, 240, and 252. This situation explains the number of test items, where the reliability coefficient increases when the sample size is 8 times the number of items up to 16 times the number of items, but does not increase which means when the number of samples is less than 8 times the number of items or greater than 16 times the number of items. The development of the instrument using 10 times the number of items has been described by Nunnally (1970, 1986) in Afif et al., (2021) and Atamimi (2014). Crocker and Algina (1986) stated that for the sake of information stability, a minimum of 200 respondents were required in the instrument trial (Kennedy, 2022; Suciati et al., 2020). Different scoring techniques affect the level of reliability of the test but do not provide a significant difference in the level of reliability for the number of samples 30 and 40, then the reliability will fluctuate according to the number of samples drawn from the population because the variation in scores is more diverse (Putri & Nahadi, 2019).

The average test reliability coefficient for all sample sizes is in the interval 0.594 – 0.658, where 0.594 is the smallest average reliability coefficient that occurs in a sample size of 216 or 18 times the number of items, while 0.658 is the largest average reliability coefficient that occurs at a sample size of 168 or 14 times the number of items. These facts provide an interesting phenomenon in that a high-reliability coefficient is not absolutely produced by the largest sample size in treatment, and vice versa, it is not necessary that a small sample size produces the lowest reliability coefficient. Relevance is stated by Shoukri et al., (2004) that the correlation between sample size and the reliability coefficient is relatively very small, the more sample size used does not necessarily result in a high reliability coefficient. In the research results, Chairunisa (2016) explained that a relatively homogeneous population will produce a small score variance so that the reliability coefficient also tends to be small. Consequently, it is not the sample size that determines the size of the reliability coefficient but the variability of the test takers' scores does.

The test reliability coefficient produced in this study, as a whole, is around 0.60 or is in the sufficient category, and is appraised to be inadequate. Some references state that a reliability coefficient smaller than 0.70 is regarded as inadequate (Ayu & Rosli, 2020; Hidayad et al., 2017; Van der Colff & Rothmann, 2009). In general, for fields of science that have high accuracy measurements such as measuring success in learning mathematics should have a high reliability coefficient that is above 0.70 (Alwi, 2015) and the good internal reliability coefficient of 0.70 or more (Taherdoost, 2016).

Judging from the characteristics of the coefficient of variation in the reliability coefficient produced by different sample sizes, this study shows fluctuations that tend to be linear in the opposite direction. That is, if the sample size is enlarged, the coefficient of variation of the reliability coefficient tends to get smaller. This indicates that the reliability coefficient of the test is stable at a larger sample size. In this study, the smallest coefficient of variation was achieved at a sample size of 204 (17 times the number of items), then at a sample size of 252 (21 times the number of items). Meanwhile, the largest coefficient of variation occurs at sample sizes of 36 and 48 (3 times and 4 times the number of items, respectively). The coefficient of variation or the dispersion coefficient shows the distribution of the normalized reliability coefficient values which are defined as the ratio of the standard deviation to the mean (Hadinata, 2018), used to compare the variability between groups (Pélabon et al., 2020). For certain purposes, if two or more coefficients of variation are compared, then the smallest coefficient of variation is the best (Alwi, 2015; Hadinata, 2018).

Referring to the results of the analysis of variance, it was found that differences in the sample size resulted in differences in the magnitude of the test reliability coefficient. In this study, the magnitude of the reliability coefficient fluctuated based on the sample size, where sample sizes of 96, 108, 120, 132, 144, 156, 168, and 192 tended to produce a higher reliability coefficient compared to sample sizes smaller and greater than 192. There is a tendency for the magnitude of the reliability coefficient of the test to approach the shape of the normal distribution relative to the sample size. The results of this study also provide information that an increase in sample size is not followed by an increase in the reliability coefficient, so there is no indication of a linear relationship. A higher reliability coefficient does not necessarily explain the characteristics of the population better, many reliability coefficients are biased upwards except in very large samples (Savalei & Reise, 2019).

The ability of test takers who are classified into high and low categories does not provide a significant difference to the reliability coefficient of the test. Reinforced by the quantitative results, the average test reliability coefficients produced by groups of high and low ability test takers were 0.629 and 0.624, respectively. Likewise, there is no significant difference in the average coefficient of variation of the reliability coefficient of the test produced between test takers with high and low abilities, each of whom has an average value of the coefficient of variation of 6.477 for high ability and 6.830 for low ability. In this study it was also found that there was no significant interaction effect between sample size and the ability of the test takers on the difference in the reliability coefficient or the difference in the coefficient of variation. Setiyawan (2014) in the results of his study explained that the group level was not very effective on test accuracy because it was equivalent to the difficulty level of the test. Very easy or very difficult tests cannot measure individual differences. Livingston (2018) stated that the test taker's ability is not consistent with changes in tests; the test taker may get a good score on test A but not good on other tests because they face different situations. Parsons et al., (2019) explained that reliability is highly dependent on random variations in scores. It is difficult to fully trust the existing set of scores because it will be different if the test is given in different situations even if it is given to the same test taker.

CONCLUSION

The conclusions of the present study are proposed as follows. First, the teacher-developed math formative test has a reliability coefficient at the interval of 0.594 – 0.658. This value is less than 0.70; so the quality of the test is categorized as inadequate. Second, the sample size does not have a linear correlation with reliability coefficients. This shows that increasing sample size does not increase reliability coefficients linearly; but, reliability coefficient changes fluctuate. Third, the reliability coefficient of the tests analysed is stable along with the enlarged sample size, meaning that the larger the sample size, the more stable the reliability coefficient will be, which is indicated by the smaller the coefficient of variation index. In this study the reliability coefficient was more stable at sample sizes of 204 (17 times the number of items) and 252 (21 times the number of items). Fourth, differences in the level of ability of the test takers do not affect the magnitude of the reliability coefficient, and there is also no interaction between the sample size and the ability of the test takers on the magnitude of the reliability coefficient. Regarding the development of test quality, further research is needed by increasing the population size, increasing the variation in the sample size relative to the number of items, and increasing the number of items. In addition, it needs to be controlled with psychological factors which are assumed to affect the accuracy of the test takers' answers, such as factors of anxiety, confidence, and thoroughness.

ACKNOWLEDGEMENTS

The researchers would like to thank the school principal and mathematics teacher at State Junior High School SMP 9, Kendari, who have assisted in the data collection process so that this research and article can be completed properly.

REFERENCES

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457

Afif, M., Suminto, A., & Mubin, A. F. (2021). Pengaruh promosi media sosial dan Word of Mouth (WOM) terhadap keputusan pembelian konsumen (studi di toko buku La Tansa Gontor) [Effects of social media promotion and words of mouth (WOM) on consusumers' decision to purchase (study in La Tansa bookstrore Gontor]. *Journal of Islamic Economics* (JoIE), *1*(2), 1–23. https://doi.org/10.21154/joie.v1i2.3206

Alwi, I. (2015). Kriteria empirik dalam menentukan ukuran sampel pada pengujian hipotesis statistika dan analisis butir [Empirical criteria in determining sample sizes in hypothesis testing statistics and item analysis]. Formatif: *Jurnal Ilmiah Pendidikan MIPA*, 2(2), 140–148. https://doi.org/10.30998/formatif.v2i2.95

Antara, I. G. W. S., Sudarma, I. K., & Dibia, I. K. (2020). The assessment instrument of mathematics learning outcomes based on HOTS toward two-dimensional geometry topic. *Indonesian Journal Of Educational Research and Review*, *3*(1), 19. https://doi.org/10.23887/ijerr.v3i2.25869

Argianti, A., & Retnawati, H. (2020). Characteristics of math national-standardized school exam test items in junior high school: What must be considered? *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(2), 156–165. https://doi.org/10.21831/pep.v24i2.32547

Ariawan, R., Zetriuslita, Z., Anggara, R. P., & Winanda, S. V. (2022). Pelatihan penyusunan soal HOTS bagi guru matematika [Training in HOTS test item writing for mathematics teachers]. *Jurnal Altifani Penelitian Dan Pengabdian Kepada Masyarakat*, *2*(1), 65–74. https://doi.org/10.25008/altifani.v2i1.207

Arif, M. (2016). Pengembangan instrumen penilaian mapel sains melalui pendekatan keterampilan proses sains SD/MI [Developing evaluation instruments for science subjects through skill process approach public/Islamic primary schools]. *Ta'allum: Jurnal Pendidikan Islam*, *4*(1), 123–148. https://doi.org/10.21274/taalum.2016.4.1.123-148

Arisandi, D., & Dewi Putri, S. (2016). SATIN-Sains dan teknologi informasi simulasi produksi gambir dengan metode supply chain management [Information technology in simulation of gambir vine product by the mngement chain supply method]. *Sains Dan Teknologi Informasi*, *2*(2), 1–8. http://jurnal.stmik-amik-riau.ac.id/index.php/satin/article/view/164

Ariyanti, E., & Bhakti, Y. B. (2020). Perbandingan bentuk tes pilihan ganda dan teknik penskoran terhadap reliabilitas tes mata pelajaran kimia [Comparison of multiple-choise test forms and scoring techniques on the reliability of the test in chemistry subject matter]. *Titian Ilmu: Jurnal Ilmiah Multi Sciences*, 12(2), 66–76. https://doi.org/10.30599/jti.v12i2.627

Arum, A. E., Khumaedi, M., & Susilaningsih, E. (2022). Validity and reliability of development of self-confidence assessment instruments for students on chemistry subject. *Journal of Research and Educational Research Evaluation*, *11*(1), 62–69. https://journal.unnes.ac.id/sju/jere/article/view/55048

Atamimi, N. (2014). Perbedaan peran jenis kelamin, skala akademik,dan peran aktif berorganisasi dengan prestasi akademik [Differences in the roles of gender, academic scale, and active roles in organization with academic achievements]. *Jurnal Cakrawala Pendidikan*, *2*(2), 236–244. https://doi.org/10.21831/cp.v2i2.2163

Ayu, S., & Rosli, M. S. Bin. (2020). Uji reliabilitas instrumen penggunaan SPADA [Reliability test on the use of SPADA instrument] (Sistem Pembelajaran Dalam Jaringan). *Biormatika*, *6*(1), 145–155. https://ejournal.unsub.ac.id/index.php/FKIP/article/view/706

Bajpai, R., & Bajpai, S. (2014). Goodness of measurement: reliability and validity. *International Journal of Medical Science and Public Health*, *3*(2), 112. https://doi.org/10.5455/ijmsph.2013.191120133

Bhakti, Y. B. (2015). Pengaruh jumlah alternatif jawaban dan teknik penskoran terhadap reliabilitas tes [Effects of number of alternatives and scoring technique on a reliability test]. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, *5*(1), 1–13. https://doi.org/10.30998/formatif.v5i1.168

Brown, G. T. L., & Harris, L. R. (2014). The future of self-assessment in classroom practice: reframing self-assessment as a core competency. *Frontline Learning Research*, *2*(1), 22–30. https://doi.org/10.14786/flr.v2i1.24

Burmeister, E., & Aitken, L. M. (2012). Sample size: How many is enough? *Australian Critical Care*, *25*(4), 271–274. https://doi.org/10.1016/j.aucc.2012.07.002

Calif, R., & Soubdhan, T. (2016). On the use of the coefficient of variation to measure spatial and temporal correlation of global solar radiation. *Renewable Energy*, 88, 192–199. https://doi.org/10.1016/j.renene.2015.10.049

Canchola, J. A. (2017). Correct use of percent coefficient of variation (%CV) formula for log-transformed data. *MOJ Proteomics & Bioinformatics*, *6*(3), 4–7. https://doi.org/10.15406/mojpb.2017.06.00200

Cassettari, L., Mosca, R., & Revetria, R. (2012). Monte Carlo simulation models evolving in replicated runs: A methodology to choose the optimal experimental sample size. *Mathematical Problems in Engineering*, 2012. https://doi.org/10.1155/2012/463873

Chairunisa, E. D. (2016). Komparasi estimasi reliabilitas pada mata pelajaran sejarah ditinjau dari homogenitas dan heterogenitas kelompok [Reliability estimation comparison on history subject matter viewed from group homogeneity and heteriogeneity]. *Jurnal Pendidikan Ilmu Sosial*, *24*(2), 179. https://doi.org/10.17509/jpis.v24i2.1454

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big dif: improved differential test functioning statistics that account for sampling variability. *Educational and Psychological Measurement*, *76*(1), 114–140. https://doi.org/10.1177/0013164415584576

Crocker, L. & Algina, J. (1986) Introduction to classical and modern test theory. Harcourt, New York

Dewi, I. P. K., Ariawan, I. P., & Gita, I. N. (2019). Analisis kesalahan pemecahan masalah matematika siswa kelas XI SMA Negeri 1 Tabanan [Error analysis on mathematics problem-solving Year XI students of State Senior Hihgh School 1 Tabanan]. *Jurnal Pendidikan Matematika Undiksha*, *10*(2), 43. https://doi.org/10.23887/jjpm.v10i2.19917

Eck, J. E., & Liu, L. (2008). Contrasting simulated and empirical experiments in crime prevention. *Journal of Experimental Criminology*, *4*(3), 195–213. https://doi.org/10.1007/s11292-008-9059-z

Gillenwater, J., Kulesza, A., Mariet, Z., & Vassilvitskii, S. (2019). A tree-based method for fast repeated sampling of determinantal point processes. *36th International Conference on Machine Learning, ICML 2019*, 2019-June, 4092–4103. https://proceedings.mlr.press/v97/gillenwater19a.html

Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., & Sodnik, J. (2014). An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. *Sensors* (*Switzerland*), *14*(2), 3702–3720. https://doi.org/10.3390/s140203702

Gunartha, I. W. (2022). Estimasi kesalahan pengukuran dalam bidang pendidikan berdasarkan teori tes klasik Estimation of mesurement error in the education field based on classical test theory]. *Jurnal Widyadari*, *23*(1), 34–47. https://doi.org/10.5281/zenodo.6390889

Hadinata, S. (2018). Tingkat pengembalian (return), risiko, dan koefisien variasi pada saham syariah dan saham nonsyariah [Return levels, risks, and variation coefficient on syariah and nonsyariah share. *AKTSAR: Jurnal Akuntansi Syariah*, *1*(2), 171. https://doi.org/10.21043/aktsar.v1i2.5089

Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, *8*. https://doi.org/10.1007/s40692-020-00169-2

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*(3), 66–67. https://doi.org/10.1136/eb-2015-102129

Herman, H., Rahim, A. R., & Syamsuri, A. S. (2021). Analisis instrumen tes hasil belajar berbasis higher order thinking skill (HOTS) [Item analysis HOTS-based learning achievement test]. *Jurnal Riset Dan Inovasi Pembelajaran*, *1*(3), 88–101. https://doi.org/10.51574/jrip.v1i3.65

Hidayad, A., Masrukan, M., & Kartono, K. (2017). Instrumen asesmen sikap siswa berbasis konservasi pada pembelajaran matematika SMP [Instrument assessment of students' attitudes based on conversion of the mathematics subject matter in the junior high school]. *Journal of Research and Educational Research Evaluation*, *6*(1), 30–38. https://journal.unnes.ac.id/sju/index.php/jere/article/view/16205

Hidayat, S. R., Setyadin, A. H., Hermawan, H., Kaniawati, I., Suhendi, E., Siahaan, P., & Samsudin, A. (2017). Pengembangan instrumen tes keterampilan pemecahan masalah pada materi getaran, gelombang, dan bunyi [Developing test instrument for problem-solving skills on the materials of vibration, wave, and sound]. *Jurnal Penelitian & Pengembangan Pendidikan Fisika*, *3*(2), 157–166. https://doi.org/10.21009/1.03206

Hikamudin, E., & Hairun, Y. (2021). Analisis disparitas skor tampak dan estimasi skor murni dengan pengkategorian acuan normatif pada tes hasil belajar siswa [Analysis of seen score disparity and estimtion of pure score with norm-referenced categorization] . *Delta-Pi: Jurnal Matematika Dan Pendidikan Matematika*, *10*(1), 138–154. https://doi.org/10.33387/dpi.v10i1.2905

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*(1), 1–15. https://doi.org/10.2165/00007256-200030010-00001

Idrus, S. W. Al. (2022). Analisis problematika evaluasi pembelajaran IPA pada masa pandemi: kajian Literatur [Problematics analysis Physics learning evaluation during the pandemic era: Literary Study]. *Jurnal Ilmiah Profesi Pendidikan*, *7*(3c), 1979–1983. https://doi.org/10.29303/jipp.v7i3c.880

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP [Analysis of TAP application-based test item quality in the university]. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Jalilibal, Z., Amiri, A., Castagliola, P., & Khoo, M. B. C. (2021). Monitoring the coefficient of variation: A literature review. *Computers and Industrial Engineering*, 161. https://doi.org/10.1016/j.cie.2021.107600

Jian, H., & Shaoqian, L. (2014). Formative assessment in L2 classroom in China: The current situation, predicament and future. *Indonesian Journal of Applied Linguistics*, *3*(2), 18–34. https://doi.org/10.17509/ijal.v3i2.266

Junika, N., Izzati, N., & Tambunan, L. R. (2020). Pengembangan soal statistika model PISA untuk melatih kemampuan literasi statistika siswa [Developing PISA-model statistics test item to train students' statistical literacy. *Mosharafa: Jurnal Pendidikan Matematika*, *9*(3), 499–510. https://doi.org/10.31980/mosharafa.v9i3.615

Jusrianto, J., Zahir, A., Nur, H., & Parubang, D. (2022). Pendampingan penyusunan analisis tes di SD Negeri 156 Wonosari [Advocating test analysis development in State Primary School 156 Wonosari]. *Abdimas Singkerru*, *2*(1), 19–22. https://doi.org/10.59563/singkerru.v2i1.123

Kapantow, B., Mananoma, T., & Sumarauw, J. S. F. (2017). Analisis debit dan tinggi muka air sungai paniki di kawasan Holland Village [Analysis of water.discharge and surface of the paniki river in Holland Village] *Jurnal Sipil Statik*, *5*(1), 21–29. https://ejournal.unsrat.ac.id/v2/index.php/jss/article/view/15734

Kartowagiran, B., & Jaedun, A. (2016). Model asesmen autentik untuk menilai hasil belajar siswa sekolah menengah pertama (SMP): implementasi asesmen autentik di SMP [Authentic assessment model to measure learning achievement of junior high school students: implementation of authentuc assessment in the junior high school]. *Jurnal Penelitian Dan Evaluasi Pendidikan*, [Research and Educational Evaluatuion Journal] *20*(2), 131–141. https://doi.org/10.21831/pep.v20i2.10063

Kasli, E., Farhan, A., Susanna, S., Herliana, F., & Wahyuni, S. (2022). Overview of teacher ability using core type cooperative model with blended learning method to increase student learning outcomes. *Jurnal Penelitian Pendidikan IPA*, *8*(2), 1012–1017. https://doi.org/10.29303/jppipa.v8i2.1241

Kennedy, I. (2022). Sample size determination in Test-Retest and Cronbach Alpha reliability estimates. *British Journal of Contemporary Education*, *2*(1), 17–29. https://doi.org/10.52589/bjce-fy266hk9

Khumaedi, M. (2012). Reliabilitas instrumen penelitian [Reliability of research instrument]. *Jurnal Pendidikan Teknik Mesin* [Mechanical Engineering Educational Journal] *Unnes*, *12*(1), 25-30. https://journal.unnes.ac.id/nju/JPTM/article/view/5273

Komperda, R., Pentecost, T. C., & Barbera, J. (2018). Moving beyond Alpha: a primer on alternative sources of single-administration reliability evidence for quantitative chemistry education research. *Journal of Chemical Education*, *95*(9), 1477–1491. https://doi.org/10.1021/acs.jchemed.8b00220

Kuh, G. D., & Ewell, P. T. (2010). The state of learning outcomes assessment in the United States. *Higher Education Management and Policy*, *22*(1), 1–20. https://doi.org/10.1787/hemp-22-5ks5dlhqbfr1

Kumar, A., & Misra, D. K. (2020). A review on the statistical methods and implementation to homogeneity assessment of certified reference materials in relation to uncertainty. *Mapan - Journal of Metrology Society of India*, *35*(3), 457–470. https://doi.org/10.1007/s12647-020-00383-4

Kummerfeld, E., & Rix, A. (2019). Simulations evaluating resampling methods for causal discovery: ensemble performance and calibration. *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM 2019, 2586–2593. https://doi.org/10.1109/BIBM47256.2019.8983327

Livingston, S. A. (2018). Test Reliability - Basic Concepts. In Research Memorandum ETS RM-18-01 (Issue January). https://www.ets.org/research/policy_research_reports/publications/report/2018/jysw.html

Magdalena, I., Hifziyah, M., Aeni, I. N., & Rahayu, R. P. (2020). Pengembangan instrumen tes siswa tingkat sekolah dasar Kabupaten Tangerang Developing test instrument for primary level students in Tangerang Regency]. *Nusantara : Jurnal Pendidik Dan Ilmu Sosial*, *2*(2), 227–237. https://ejournal.stitpn.ac.id/index.php/nusantara/article/view/808

Makhrus, M. (2018). Analisis rencana pelaksanaan pembelajaran (RPP) terhadap kesiapan guru sebagai "Role Model" keterampilan abad 21 pada pembelajaran IPA SMP [Analysis of teacher lesson plan on teacher's readiness as 21st-skill role model in junior high school physics learning]. *Jurnal Penelitian Pendidikan IPA*, *5*(1). https://doi.org/10.29303/jppipa.v5i1.171

Mudanta, K. A., Astawan, I. G., & Jayanta, I. N. L. (2020). Instrumen penilaian motivasi belajar dan hasil belajar IPA siswa kelas V sekolah dasar [Evaluation instrument learning motivation and learning achievement Grade V Physics primary school]. *Mimbar Ilmu*, *25*(2), 101. https://doi.org/10.23887/mi.v25i2.26611

Muluki, A. (2020). Analisis kualitas butir tes semester ganjil mata pelajaran IPA kelas IV MI Radhiatul Adawiyah [Analysis of test item quality Physics subject matter Grade IV odd semester Islamic primary school]. *Jurnal Ilmiah Sekolah Dasar*, *4*(1), 86. https://doi.org/10.23887/jisd.v4i1.23335

Mustopa, A., Jasim, J., Basri, H., & Barlian, U. C. (2021). Analisis standar penilaian pendidikan [Analysis of Educational Evaluation standard]. *Jurnal Manajemen Pendidikan*, *9*(1), 24–29. https://doi.org/10.33751/jmp.v9i1.3364

Ndiung, S., & Jediut, M. (2020). Pengembangan instrumen tes hasil belajar matematika peserta didik sekolah dasar berorientasi pada berpikir tingkat tinggi [Developing HOTS-oriented mathematics learning result test for students of the primary school]. *Premiere Educandum: Jurnal Pendidikan Dasar Dan Pembelajaran*, *10*(1), 94. https://doi.org/10.25273/pe.v10i1.6274

Nopriyeni, Prasetyo, Z. K., & Djukr. (2019). The implementation of mentoring based learning to improve pedagogical knowledge of prospective teachers. *International Journal of Instruction*, *12*(3), 529–540. https://doi.org/10.29333/iji.2019.12332a

Nunnally, J. C., Jr. (1970). Introduction to psychological measurement. McGraw-Hill.

Nunnally, J. C., Jr. (1978). Psychometric theory. 2nd Edition. McGraw-Hill. New York.

Nuriyah, N. (2014). Evaluasi pembelajaran: Sebuah kajian teori [Learning evaluatrion: a theoretical analysis]. *Jurnal Edueksos*, *3*(1), 73–86. https://doi.org/10.1165/rcmb.2013-0411OC

Oktadini, N. R., Sevtiyuni, P. E., & Bardadi, A. (2022). Pelatihan aplikasi pengolah nilai rapor berbasis komputer pada guru di SMP Negeri 58 Palembang [Training of computer-based Grade-report management application teachers of State Junior High School 58 Palembang]. *Bulletin of Community Service in Information System (BECERIS)*, *1*(1), 7–13. https://doi.org/10.36706/beceris.v1i1.2

Ono, S. (2020). Uji validitas dan reliabilitas alat ukur SG Posture Evaluation [Validity and reliability test SG Posture Evaluation instrument]. *Jurnal Keterapian Fisik*, *5*(1), 55–61. https://doi.org/10.37341/jkf.v5i1.167

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821–846. https://doi.org/10.1002/tea.21316

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pélabon, C., Hilde, C. H., Einum, S., & Gamelon, M. (2020). On the use of the coefficient of variation to quantify and compare trait variation. *Evolution Letters*, *4*(3), 180–188. https://doi.org/10.1002/evl3.171

Phillips, J. J., & Phillips, P. P. (2016). *Handbook of training evaluation and measurement methods, fourth edition. in handbook of training evaluation and measurement methods, Fourth Edition*. Routledge. https://doi.org/10.4324/9781315757230

Primasari, I. F. N. D., Marini, A., & Sumantri, M. S. (2021). Analisis kebijakan dan pengelolaan pendidikan terkait standar penilaian di sekolah dasar [Analysis of educational policy and management releted to evaluation standard primary school]. *Jurnal Basicedu*, *5*(3), 1479–1491. https://doi.org/10.31004/basicedu.v5i3.956

Putri, D., & Nahadi. (2019). Perbandingan reliabilitas tes hasil belajar matematika SMA berdasarkan teknik penskoran dan ukuran sampel [Comparison of reliabilty tests mathematics learning achievement senior high school based on scoring techniques and sample sizes]. *Journal Education and Chemistry (JEDCHEM)*, *1*(1), 10–24. https://doi.org/https://dx.doi.org/10.36378/jedchem.v1i1.86

Rapono, M., Safrial, S., & Wijaya, C. (2019). Urgensi penyusunan tes hasil belajar: Upaya menemukan formulasi tes yang baik dan benar Urgencies of learning achievement test construction: Efforts finding correct and good test formulations]. *Jupiis: Jurnal Pendidikan Ilmu-Ilmu Sosial*, *11*(1), 95. https://doi.org/10.24114/jupiis.v11i1.12227

Reiter-Palmon, R., Forthmann, B., & Barbot, B. (2019). Scoring divergent thinking tests: A review and systematic framework. *Psychology of Aesthetics, Creativity, and the Arts*, *13*(2), 144–152. https://doi.org/10.1037/aca0000227

Retnawati, H., Hadi, S., & Nugraha, A. C. (2016). Vocational high school teachers' difficulties in implementing the assessment in curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction*, *9*(1), 33–48. https://doi.org/10.12973/iji.2016.914a

Sarwanto, Fajari, L. E. W., & Chumdari. (2020). Open-Ended questions to assess critical-thinking skills in indonesian elementary school. *International Journal of Instruction*, *14*(1), 615–630. https://doi.org/10.29333/IJI.2021.14137A

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology*, *5*(1), 1–8. https://doi.org/10.1525/collabra.247

Schiel, J. E., Turner, A., Mouchahoir, T., Yandrofski, K., Telikepalli, S., King, J., DeRose, P., Ripple, D., & Phinney, K. (2018). The NISTmAb reference material 8671 value assignment, homogeneity, and stability. *Analytical and Bioanalytical Chemistry*, *410*(8), 2127–2139. https://doi.org/10.1007/s00216-017-0800-1

Setiyawan, A. (2014). Faktor-faktor yang mempengaruhi reliabilitas tes [Factors affecting test reliability]. *Jurnal An Nûr*, *6*(2), 341–354. https://jurnalannur.ac.id/index.php/An-Nur/article/view/53

Shavelson, R. J., Zlatkin-Troitschanskaia, O., & Mariño, J. P. (2018). *International performance assessment of learning in higher education (iPAL): Research and Development*. Assessment of Learning Outcomes in Higher Education, March, 193–214. https://doi.org/10.1007/978-3-319-74338-7_10

Shoukri, M. M., Asyali, M. H., & Donner, A. (2004). Sample size requirements for the design of reliability study: Review and new results. *Statistical Methods in Medical Research*, *13*(4), 251–271. https://doi.org/10.1191/0962280204sm365ra

Sianturi, R. (2022). Uji homogenitas sebagai syarat pengujian analisis [Test of homogeneity as a requirement for testing analysis]. *Jurnal Pendidikan, Sains Sosial, Dan Agama*, *8*(1), 386–397. https://doi.org/10.53565/pssa.v8i1.507

Sinaga, N. A. (2016). Pengembangan tes kemampuan pemecahan masalah dan penalaran matematika siswa SMP kelas VIII [Developing problem-solving skill test and mathematics reasoning students of Year VIII junior high school]. *PYTHAGORAS: Jurnal Pendidikan Matematika*, *11*(2), 169. https://doi.org/10.21831/pg.v11i2.10642

Singh, R., & Sarkar, S. (2015). Does teaching quality matter? Students learning outcome related to teaching quality in public and private primary schools in India. *International Journal of Educational Development*, 41, 153–163. https://doi.org/10.1016/j.ijedudev.2015.02.009

Suardipa, I. P., & Primayana, K. H. (2020). Peran desain evaluasi pembelajaran untuk meningkatkan kualitas pembelajaran [Roles of learning evaluation design to improve quality of instruction]. *Widyacarya*, *4*(2), 88–100. https://doi.org/https://doi.org/10.55115/widyacarya.v4i2.796

Suchyadi, Y., Sundari, F. S., Sutisna, E., Sunardi, O., Budiana, S., Sukmanasa, E., & Windiyani, T. (2020). Improving the ability of elementary school teachers through the development of competency based assessment instruments in teacher working group, north bogor city. *Journal of Community Engagement*, *2*(1), 1–5. https://journal.unpak.ac.id/index.php/jce/article/view/2742

Suciati, S., Munadi, S., Sugiman, S., & Febriyanti, W. D. R. (2020). Design and validation of mathematical literacy instruments for assessment for learning in Indonesia. *European Journal of Educational Research*, *9*(2), 865–875. https://doi.org/10.12973/eu-jer.9.2.865

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *SSRN Electronic Journal*, *5*(3), 28–36. https://doi.org/10.2139/ssrn.3205040

Umami, R., Rusdi, M., & Kamid, K. (2021). Pengembangan instrumen tes untuk mengukur higher order thinking skills (HOTS) berorientasi programme for international student asessment (PISA) pada peserta didik [Developing test instrument to measure higher order thinking skills (HOTS) oriented to programme for international student asessment (PISA). *JP3M (Jurnal Penelitian Pendidikan Dan Pengajaran Matematika)*, *7*(1), 57–68. https://doi.org/10.37058/ jp3m.v7i1.2069

Utami, R. F., Prasetyo, S., & Nuridzin, D. Z. (2022). Validitas dan reliabilitas kuesioner Chinese Positive Youth Development Scales (CPYDS) mengukur keterampilan hidup pelajar SMP di Babakan Madang Kabupaten Bogor 2019 [Validity and reliability of questionnaire Chinese Positive Youth Development Scales (CPYDS) measuring life skills of junior high school students in Babakan Madang Bogor Rrgrncy]. *Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan*, *2*(3), 125. https://doi.org/10.51181/bikfokes.v2i3.6082

Van der Colff, J. J., & Rothmann, S. (2009). Occupational stress, sense of coherence, coping, burnout and work engagement of registered nurses in South Africa. *SA Journal of Industrial Psychology*, *35*(1), 1–10. https://doi.org/10.4102/sajip.v35i1.423

Yuniartik, H., Hidayah, T., & Nasuka. (2017). Evaluasi pembelajaran pendidikan jasmani olahraga dan kesehatan di SLB C se-kota Yogyakarta [Evaluation learning physical education sport and health C Special Schools through out Yogyakarta City]. *Journal of Physical Education and Sports*, *6*(2), 148–156. https://journal.unnes.ac.id/sju/jpes/article/view/17389

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Pant, H. A. (2018). *Assessment of learning outcomes in higher education. Handbook on Measurement, Assessment, and Evaluation in Higher Education (2nd Edition)*. Routledge. https://doi.org/10.4324/9781315709307-54