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

Busnawir 1*, Kodirun 2, Zamsir 3, Hafiuddin Samparaja 4, Hasnawati 5
Department of Mathematics Education, Universitas Halu Oleo, Kendari, Indonesia

* Corresponding Author. E-mail: busnawir@uho.ac.id

ARTICLE INFO

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:

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; Oktdini 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 outcomes tests developed by teachers in schools still needs to be studied (Ndiung & Jeduit, 2020; Singh & Sarkar, 2015). There are still many teachers who do not understand in compiling item grids...
A) 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 (Suwardi & 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; Jalilbal 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; Nuriah, 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 & Atken, 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 Aff 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. Argjanti & 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 (Arisand & 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>
<thead>
<tr>
<th>Test Taker’s Ability</th>
<th>Repetition to</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>S-36</td>
<td>r_{1.1T}</td>
<td>r_{1.2T}</td>
</tr>
<tr>
<td>S-48</td>
<td>r_{2.1T}</td>
<td>r_{2.2T}</td>
</tr>
<tr>
<td>S-60</td>
<td>r_{i.1T}</td>
<td>r_{i.2T}</td>
</tr>
<tr>
<td>S-72</td>
<td>r_{31.1T}</td>
<td>r_{31.2T}</td>
</tr>
<tr>
<td>S-84</td>
<td>r_{11.1T}</td>
<td>r_{11.2T}</td>
</tr>
<tr>
<td>S-96</td>
<td>r_{12.1T}</td>
<td>r_{12.2T}</td>
</tr>
<tr>
<td>S-108</td>
<td>r_{13.1T}</td>
<td>r_{13.2T}</td>
</tr>
<tr>
<td>S-120</td>
<td>r_{14.1T}</td>
<td>r_{14.2T}</td>
</tr>
<tr>
<td>S-132</td>
<td>r_{15.1T}</td>
<td>r_{15.2T}</td>
</tr>
<tr>
<td>S-144</td>
<td>r_{16.1T}</td>
<td>r_{16.2T}</td>
</tr>
<tr>
<td>S-156</td>
<td>r_{17.1T}</td>
<td>r_{17.2T}</td>
</tr>
<tr>
<td>S-168</td>
<td>r_{18.1T}</td>
<td>r_{18.2T}</td>
</tr>
<tr>
<td>S-180</td>
<td>r_{19.1T}</td>
<td>r_{19.2T}</td>
</tr>
</tbody>
</table>

Description:
- r_{1.1T} = reliability coefficient in the 1st repetition, 1st sample, high ability category.
- r_{31.19T} = reliability coefficient in the 31st repetition, 19th sample, high ability category.
- r_{11.1R} = reliability coefficient in the 1st repetition, 19th sample, high ability category.
- r_{31.19R} = 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 α = 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. (Jalilbal 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-exploratory 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

<table>
<thead>
<tr>
<th>No.</th>
<th>Sample Size</th>
<th>N</th>
<th>Mean</th>
<th>SE Mean</th>
<th>95% Confidence Intervals</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S-36</td>
<td>31</td>
<td>0.599</td>
<td>0.014</td>
<td>(0.570797; 0.626622)</td>
<td>12.71</td>
</tr>
<tr>
<td>2</td>
<td>S-48</td>
<td>31</td>
<td>0.622</td>
<td>0.014</td>
<td>(0.593964; 0.649455)</td>
<td>12.17</td>
</tr>
<tr>
<td>3</td>
<td>S-60</td>
<td>31</td>
<td>0.596</td>
<td>0.009</td>
<td>(0.576327; 0.614835)</td>
<td>8.81</td>
</tr>
<tr>
<td>4</td>
<td>S-72</td>
<td>31</td>
<td>0.616</td>
<td>0.008</td>
<td>(0.599985; 0.631241)</td>
<td>6.92</td>
</tr>
<tr>
<td>5</td>
<td>S-84</td>
<td>31</td>
<td>0.608</td>
<td>0.009</td>
<td>(0.589169; 0.625863)</td>
<td>8.23</td>
</tr>
<tr>
<td>6</td>
<td>S-96</td>
<td>31</td>
<td>0.655</td>
<td>0.008</td>
<td>(0.638696; 0.670724)</td>
<td>6.67</td>
</tr>
<tr>
<td>7</td>
<td>S-108</td>
<td>31</td>
<td>0.645</td>
<td>0.009</td>
<td>(0.626333; 0.663602)</td>
<td>7.88</td>
</tr>
<tr>
<td>8</td>
<td>S-120</td>
<td>31</td>
<td>0.654</td>
<td>0.008</td>
<td>(0.637251; 0.671007)</td>
<td>7.03</td>
</tr>
<tr>
<td>9</td>
<td>S-132</td>
<td>31</td>
<td>0.639</td>
<td>0.008</td>
<td>(0.622759; 0.656080)</td>
<td>7.10</td>
</tr>
<tr>
<td>10</td>
<td>S-144</td>
<td>31</td>
<td>0.651</td>
<td>0.007</td>
<td>(0.636907; 0.665286)</td>
<td>5.94</td>
</tr>
<tr>
<td>11</td>
<td>S-156</td>
<td>31</td>
<td>0.642</td>
<td>0.007</td>
<td>(0.627049; 0.656628)</td>
<td>6.28</td>
</tr>
<tr>
<td>12</td>
<td>S-168</td>
<td>31</td>
<td>0.658</td>
<td>0.006</td>
<td>(0.645433; 0.670116)</td>
<td>5.12</td>
</tr>
<tr>
<td>13</td>
<td>S-180</td>
<td>31</td>
<td>0.657</td>
<td>0.006</td>
<td>(0.644844; 0.669220)</td>
<td>5.06</td>
</tr>
<tr>
<td>14</td>
<td>S-192</td>
<td>31</td>
<td>0.642</td>
<td>0.006</td>
<td>(0.628801; 0.654489)</td>
<td>5.46</td>
</tr>
<tr>
<td>15</td>
<td>S-204</td>
<td>31</td>
<td>0.610</td>
<td>0.004</td>
<td>(0.602510; 0.618070)</td>
<td>3.48</td>
</tr>
<tr>
<td>16</td>
<td>S-216</td>
<td>31</td>
<td>0.594</td>
<td>0.006</td>
<td>(0.582009; 0.606443)</td>
<td>5.61</td>
</tr>
<tr>
<td>17</td>
<td>S-228</td>
<td>31</td>
<td>0.601</td>
<td>0.005</td>
<td>(0.589782; 0.612153)</td>
<td>5.07</td>
</tr>
<tr>
<td>18</td>
<td>S-240</td>
<td>31</td>
<td>0.608</td>
<td>0.005</td>
<td>(0.598677; 0.617581)</td>
<td>4.24</td>
</tr>
<tr>
<td>19</td>
<td>S-256</td>
<td>31</td>
<td>0.597</td>
<td>0.004</td>
<td>(0.589292; 0.604902)</td>
<td>3.56</td>
</tr>
</tbody>
</table>

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
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).

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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>18</td>
<td>0.317</td>
<td>0.018</td>
<td>8.950</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>551</td>
<td>1.084</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>569</td>
<td>1.401</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S = 0.044</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-Sq = 22.63%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students Ability (AoTT)</td>
<td>1</td>
<td>0.004</td>
<td>0.004</td>
<td>1.950</td>
<td>0.180</td>
</tr>
<tr>
<td>Sample Size</td>
<td>18</td>
<td>0.317</td>
<td>0.018</td>
<td>9.260</td>
<td>0.000</td>
</tr>
<tr>
<td>Students Ability*Sample Size</td>
<td>18</td>
<td>0.034</td>
<td>0.002</td>
<td>0.970</td>
<td>0.496</td>
</tr>
<tr>
<td>Error</td>
<td>532</td>
<td>1.046</td>
<td>0.002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>569</td>
<td>1.401</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

![Boxplot of high and low AoTT variation coefficients](image)

**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


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 heterogeneity]. Jurnal Pendidikan Ilmu Sosial, 24(2), 179. https://doi.org/10.17509/jpis.v24i2.1454


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

