REVIEW OF INTERNATIONAL GEOGRAPHICAL EDUCATION

ISSN: 2146-0353 ● © RIGEO ● 11(5), SPRING, 2021

www.rigeo.org

Research Article

Retest on Online Test: How Stable and Reliable?

Bulkani[1]
Muhammadiyah University of Palangkaraya
bulkaniardiansyah@gmail.com

Rita Rahmaniati[2]
Muhammadiyah University of Palangkaraya

Chandra Anugrah Putra[3]
Muhammadiyah University of Palangkaraya

Abstract

The study aims to determine (1) the stability of online test scores if used repeatedly, (2) changes in reliability or internal consistency in online tests if used repeatedly, and (3) the effect of the number of test-takers and the number of online test items on the reliability of the online test. The research uses a quantitative approach in November 2020-January 2021, involving 134 students of the Muhammadiyah University of Palangka Raya who took four different subjects. The research procedure is (1) Arranging several test kits on various topics and the number of items, (2) Carrying out theoretical validity tests by asking for reviews from several relevant experts, (3) Assigning test subjects, (4) Carrying out online tests with techniques test-retest. With this technique, a group of test-takers is given the same or unidimensional set of tests several times. The test contents' differences in the 1st, 2nd, and 3rd tests are only in the items' serial number. The period between implementing one test and another test is two weeks, (5) Analyzing the test results to see changes in the test's reliability on several repetitions. Data analysis is performed using ANATES software version 409/2004. The results showed (1) The online test score used for three repetitions is relatively stable. At the 5% significance level, there is a significant positive correlation between test scores on the 1st, 2nd, and 3rd repetitions. It proves that the test result score is stable between repetitions, (2) The reliability or internal consistency of online tests tends to show an increase if used repeatedly. The online test's reliability reached a sufficient number in the 3rd repetition, (3) Descriptively, the number of test items or the test's length affects the online test's reliability. The more the number of items, the higher the reliability of the online test. Also, the number of test-takers does not affect the reliability of online tests. The research implies that it is advisable to repeat the test-retest technique three times to obtain adequate online test results. Using unidimensional online test items and changing the test serial number is also recommended to reduce the carry-over effect.

Keywords

To cite this article: Bulkani, Rahmaniati R, and Putra C, A. (2021). Retest on Online Test: How Stable and Reliable? Review of International Geographical Education (RIGEO), 11(5), 2581-2590 Doi: 10.48047/rigeo.11.05.155

Submitted: 02-11-2020 ● **Revised:** 15-02-2021 ● **Accepted:** 25-03-2021

Introduction

The Covid-19 pandemic that has hit the world has changed the order of life. The pandemic has changed the learning system in education. The learning system, which is implemented initially offline, is forcibly changed to online implementation. The sudden change from offline learning to online learning leaves many problems. Online learning, among others, tends to make students confused, bored, passive, less creative, less productive, stressful due to piling up homework (Argaheni, 2020). Students experience mild anxiety (Hasanah et al., 2020), suffering psychologically, and isolated from teachers and friends' interactions (Pietro et al., 2020). The positive impacts include an increase in literacy skills (Argaheni, 2020), an increase in learning independence and an increase in the ability to use information technology (Firman & Rahayu, 2020), and the growth of healthier living habits (Sukandi et al., 2021). On the other hand, many teachers are not ready to carry out online learning. Most of the teachers' online knowledge is only through giving assignments because only 67% of teachers can use digital devices (Rahiem, 2020). In Indonesia, the Ministry of Education and Culture has implemented an emergency curriculum based on online learning. The Ministry of Education and Culture of the Republic of Indonesia issued circular letter number 4 of 2020 concerning the Implementation of Education Policies in an Emergency for the Spread of Corona Virus Disease (Kemdikbud, 2020). Some of the critical points are about the procedures and adjustments to implementing learning and guidelines for evaluating learning outcomes, such as national and school exams during the pandemic. These changes also impact learning orientation and evaluation techniques, which can accommodate each area's characteristics and the different situations each student faces. For example, differences in internet signals' quality and stability can cause delays in students responding to online tests and assignments given by teachers. It can lead to differences in the learning outcomes they get. Also, evaluating learning outcomes that are not directly supervised has decreased the evaluation results' accuracy. For this reason, many parties have suggested that education policymakers should immediately prepare a learning outcome assessment guide to reduce weaknesses in the online evaluation. There must be an appropriate strategy to ensure a more accurate learning outcome evaluation system. The focus must take into account the pre-conditions before the online assessment is carried out by taking into account the differences in situations faced by students (Jimenez, 2020), the format and the right time for conducting the evaluation, and the possibility of decreasing the Validity and reliability of the instruments used (Rahim, 2020). Online learning also forces a change in the orientation of learning outcomes assessment to be solely on learning outcomes. The essence of education that prioritizes the teacher cannot observe the process. In other words, teachers only receive learning outcomes products without directly supervising the learning process and working on assignments and test questions. It is the possibility of refraction, thereby reducing the level of confidence in the assessment results. Diningrat, Nindya, and Salwa (2020) state that one of the barriers to online learning is the low reliability of learning evaluation results. One of the instruments for assessing online learning outcomes is an online test. The intended use, substance, and online test materials are the same as offline tests, but the difference between the two lies in the way they are used. Online examinations are given with a specific time limit, for example, using the google-form format. Some of the advantages of online testing are its broad coverage and space, unlimited number of test-takers to work on simultaneously, and the ease of giving test scores because a computer system automatically carries it out. On the other hand, the weaknesses are the influence of internet signal quality on the speed of sending test results and the lack of supervision in its operation. It will reduce the level of confidence of the meter against the test results. Also, the quality and accuracy of the online test results are influenced by the test items' consistency and characteristics. The more consistent the measurement results are, the higher the reliability of the test. A test score is reliable if it tends to be consistent on the same or different test editions but measures the same ability and shows consistency when used by different gauges (Livingston et al., 2018). Thus, the test's reliability is reflected in the test results' consistency, which can be seen from the changes in the test items' characteristics. The consistency of a test's measurement results can be seen by taking repeated measurements, if these measurements can be made using several unidimensional test kits, or using the same test (test-retest). Repeated use of online tests can increase reliability (Rapanta et al., 2020). Naga stated that, theoretically, the error score on the measurement results tends to approach zero if an infinite number of measurements are taken

(1992). Meanwhile, Azwar (1987) argues that the various limitations that the test has as a means of measuring learning outcomes can be overcome by continuous measurement. Analysis of repeated tests' reliability will help many parties use the test more widely (Aldridge, Dovey, & Wade, 2017). Reliable test kits and test items function optimally as an accurate measuring tool to describe students' actual abilities. It is achieved if the situation at the time of the test can be controlled. In other words, the evaluator has confidence that the test is done independently by the test-taker. However, in the use of online tests, less than optimal supervision in implementing the test can reduce confidence. Therefore, several strategies are needed to improve the accuracy of online test results. Livingston et al. (2018) suggests the importance of a strategy to control sources of inconsistencies such as evaluator selection, timing, test day, and specific selection of test items. While Aldridge et al. (2017) state that the use of the same or equivalent test repeatedly, also known as the test-retest technique, is needed in measuring psychomotor aspects to determine individual differences regardless of time context and other factors. Spencer (2003) and Heise (1969) suggest using the term stability to describe the same test's reliability, and it is used repeatedly. In contrast, reliability refers more to the internal correlation on a test. Berchtold (2016) emphasizes the use of the term reliability as a correlation between the score of a measurement result with other equivalent measurement results, which is also significant for the test results' stability. In using online tests, so far, most educators only use them with one measurement. The changes are only in the way it is used, from offline to online. On the other hand, taking online measures an infinite number of times is impossible (Naga, Geiger, & Muller, 1992). Repetition of the test as many times as possible will take a lot of time, cost, and effort, making it impossible to carry out. For this reason, it is necessary to observe changes in the level of reliability or stability of online test items so that the minimum number of repetitions can be determined so that the measurement results are sufficiently consistent and stable. Some of the research problems that must be answered: (1) Are the online test scores, if repeatedly used, stable? (2) How is the reliability or internal consistency of online tests if used repeatedly? Online test on online test reliability? The research is essential to provide educators with an overview of online tests' reliability or consistency if used repeatedly. Educators can use the research results to reference one way of using online tests to obtain more accurate results. Educators as evaluators need to understand some of the risks of using online tests and how to reduce the risks so that measurement results are obtained that are pretty reliable and describe students' actual abilities.

Research Method

The study uses a quantitative approach. The research time was November 2020-January 2021. The research involves 134 students of the Muhammadiyah University of Palangka Raya who take four different subjects, as shown in the following table:

Table 1:
Types of courses and the number of online test-takers

No	Courses	Number of test-takers
1.	Evaluation of Primary School Learning	34
2.	Quantitative Research Methodology	50
3	Communication Skills	26
4.	Basic Natural Sciences	24

The research hypotheses tested are a significant positive correlation between online test scores on the 1st repetition and online test scores on the 2nd and 3rd repetitions. The research procedure is (1) Arranging several test kits on different subjects and several items, (2) Carrying out theoretical validity tests by asking for reviews from several relevant experts, (3) Assigning test subjects, (4) Carrying out online tests with techniques test-retest. With this technique, a group of test-takers is given the same set of tests several times. The test contents' differences in the 1st, 2nd, and 3rd tests were only in the items' serial number. The time interval between implementing one test and another test is two weeks, (5) Analyzing the test results to see changes in the test's reliability on several repetitions. Analysis of test scores using ANATES software version 409/2004, primarily to

calculate the reliability or internal consistency of online tests, (6) Perform descriptive and inferential data analysis. Researchers use SPSS version 2.0 software for correlation testing. All statistical tests use a significance level of 5%.

Results and Discussions

Correlation between test results on three repetitions

The study also finds a significant correlation between online test results on tests 1, 2, and 3, as illustrated in the following table.:

Table 2:
Correlation of online test results with three repetitions

Courses / Repetition to		2	3
	1	0.68	0.58
Evaluation of Primary School Learning	2		0.73
	1	0.54	0.49
Research Methodology. Quantitative	2		0.55
	1	0.87	0.77
Communication Skills	2		0.75
	1	0.53	0.49
Basic Natural Sciences	2		0.84

The table above shows a significant positive correlation between the 1st, 2nd, and 3rd repeat online tests at the 5% significance level and degrees of freedom (34-1). The correlation is included in the medium and high categories. The significant correlation is evidence that the online test scores on the 1st, 2nd, and 3rd iterations are consistent. Consistency is an important indicator to describe the reliability of the test.

Internal reliability / consistency changes in 3 repetitions

From the research results, it is obtained an overview of changes in test reliability as follows:

Table 3:
Reliability of the online test with 3 repetitions

No	Courses	1st	2nd	3rd	Average
1.	Evaluation of Primary School Learning	0.25	0.52	0.71	0.49
2.	Quantitative Research Methodology	0.52	0.66	0.67	0.62
3.	Communication Skills	0.64	0.75	0.78	0.72
4.	Basic Natural Sciences	0.77	0.89	0.92	0.86
	Average	0.55	0.71	0.77	-

Test reliability has increased from the 1st, 2nd, and 3rd repetitions. It means that the online test tends to be more reliable if it is given repeatedly. In the first iteration, the online test's reliability tended to be relatively low, but the increase in reliability occurred in the 2nd and 3rd repetitions. In the 3rd repetition, in general, the online test's reliability had reached a sufficient number, which was above 0.67. The average reliability has also increased, judging from the results of online tests on all courses. In the 1st repetition, the average reliability is 0.55. In the second repetition, the average reliability is 0.71, and the 3rd iteration has obtained an average of 0.77. It indicates that the reliability of the online test used has increased with the number of repetitions.

Test length and reliability

The data is obtained judging from the length of the online test or the number of items used:

Table 4:
Test Length and Average Reliability

No	Courses	Number of Test Items	Average Reliability
1.	Evaluation of Primary School Learning	30	0.49
2.	Quantitative Research Methodology	30	0.62
3.	Communication Skills	40	0.72
4.	Basic Natural Sciences	50	0.86

Descriptively, the table above shows that increasing the number of items on the online test increases the test's average reliability. In a subject with 30 test items, the average reliability on three repetitions is 0.49 and 0.62. In the online test with 40 test items, average reliability of 0.72 is obtained. Whereas in the online test with the number of items 50, the test's average reliability on three repetitions is 0.86.

Number of test-takers and reliability

When viewed from the number of test-takers, the reliability average is obtained as follows :

Table 5:
Number of Test Participants and Average Reliability

No	Courses	Number of Test-Takers	Average Reliability
1.	Evaluation of Primary School Learning	34	0.49
2.	Quantitative Research Methodology	50	0.62
3.	Communication Skills	26	0.72
5.	Basic Natural Sciences	24	0.86

The table above shows that, descriptively, the number of online test-takers is not related to the average test reliability at three repetitions. It means that an increase in test-takers number does not necessarily increase the online test's average reliability and vice versa.

Discussions

The study's results indicate that the correlation between the 1st, 2nd, and 3rd tests is significant. It proves that the score of the repeated test results is relatively stable. The stability of the online test, which is used repeatedly, can be seen from the correlation coefficient between the results of these multiple measurements. Wells and Wollack (2003) state that the test-retest reliability is the consistency of the measured score, even though the test taker did not produce the same score on several measurements. Thus, the correlation coefficient between the results of one test and subsequent tests using the same or equivalent test can describe the test's reliability. The type of reliability is often referred to as a stability test. The degree of correlation between test scores shows the test's degree of reliability (Hajjar, 2018). Wells and Wollack (2003) also emphasize the importance of equality between first and subsequent test results. Aldridge et al. (2017) state the importance of identifying the difference between the first and next test scores to find repeatedly used tests' reliability or stability. Good measurement results show consistency between times, raters, and unidimensional test kits. The study finds a significant correlation coefficient between test scores on the 1st, 2nd, and 3rd repetitions. It is an indication that the online test used is relatively stable and reliable if used repeatedly. The results also indicate that repetition of the online test

would result in a more reliable test score. Based on the description of the correlation coefficient in Table 5, r23> r12, and r23> r13. The correlation coefficient on the 2nd and 3rd repetition test scores results is the highest compared to the 1st and 2nd repetitions. The phenomenon is an indication that the online test results tested tend to be stable and quite adequate on the 3rd iteration. Suppose it is related to the Spearman-Brown prediction formula by assuming that the 1st, 2nd, and 3rd repetitions are part of the unidimensional test. In that case, the phenomenon of the correlation between the test results can be described as follows:

$$\rho_{ij} = \frac{2r_{ij}}{(1 + r_{ij})} \quad \text{..... (Finch \& French, 2018; Naga et al., 1992; Widoyoko, 2011)}$$

di mana :
ρ_{ij} = Spearman-Brown Reliability
r_{ij} = the correlation coefficient between the ith and jth tests

The formula above shows that the increase in Spearman-Brown reliability will be in line with the correlation coefficient's rise between the repetition results. The higher the correlation coefficient between the two test results, the Spearman-Brown reliability also increases. In general, it can be seen that the values of r23> r12, and r23> r13 are as illustrated in Table 5, so that ρ23> ρ12, and ρ23> ρ13. In other words, the Spearman-Brown reliability will also increase in line with the number of test repetitions. In terms of internal consistency, the study proves that the online test's reliability coefficient repeatedly tends to increase. Although the reliability coefficient is not sufficient in the first test, the reliability coefficient is adequate, above 0.60, on the 3rd iteration. Taherdoost (2016) suggests the reliability coefficient must be above 0.60. While Hinton et al. (2004) classify high-reliability coefficients if above 0.70, medium reliability is in the range of 0.50-0.70 and low if it is below 0.50. Naga et al. (1992) states that the reliability classification depends on the type and purpose of measurement, but 0.60 is generally considered sufficient. Reliability test describes the consistency of measurement results. An overview of reliability is needed to clarify the accuracy of measurement results, especially in indirect measurements where the measurer does not directly find an accurate measurement result score. In indirect measurement, the evaluator cannot directly determine the size of the object. The evaluator can only measure the symptoms caused by these objects (Naga et al., 1992). An example of indirect measurement is the measurement of student learning outcomes using a set of tests. The tool used in the form of the test can only measure students' symptoms; it does not measure learning outcomes directly. The indirect measurement score consists of a primary score component and an error score component through the equation X = T + ε (Sax & Reiter, 1980). Thus, in an indirect measurement, the measurement result score X is always assumed to contain a component of the actual score and an error or error score. If someone gets a score of 50 from a test, it does not necessarily reflect the test taker's true ability because there is a possibility of an error or ε. Theoretically, there is an infinite number of possible combinations of the sum of scores T and E to produce a score of X, for example:

X = T + ε
50 = 40 + 10
50 = 60 – 10
50 = 35 + 15, etc

A test taker's ability to particular material should be constant even though there are infinitely many possible pairs of scores for X, T, and ε. A test taker's actual ability should be constant if measured using the same test or different tests, as long as the test measures the same (unidimensional) aspects and indicators. The test taker's actual ability to a test material should also be constant if measured at a different but unidimensional time, place, condition, and measuring instrument. Thus, in the equation X = T + ε, when we measure a test taker's ability using a set of tests, there is a tendency to apply Xi = T + εi. It means that even until the ith measurement, the T score will not change. Because T's value is constant, the change in the score measured by Xi is highly dependent on the difference in the value εi. On the other hand, the value of εi is random, with the mean close to zero (Naga et al., 1992). If an infinite number of repetitions are carried out, or i → ∞, the mean value of the error distribution is close to zero (με≈0, for I → ∞). Thus, in the situation i → ∞ applies:

Xi = T + εi, for i ➔ ∞, then με≈0, so μX≈ μT. Since T is constant, μT = T, so μX≈T applies. The value of μX≈T means that in infinite repetitions of the test, the X measurement results' mean score will be close to the T score, which is the test taker's true ability.

Theoretically, the repetition of as many tests as possible will result in a more accurate and reliable measurement score. In this context, reliability (ρ) is defined as the comparison between the variance of the measured score (δ2X) to the variance of the actual score (δ2T) or expressed in the equation ρ = δ2X / δ2T (Naga et al., 1992; Sax & Reiter, 1980). So, the closer the value of 2X to the value of δ2T, the reliability value of ρ is closer to 1.00. In infinite repetitions (i ➔∞), the mean score of X will be close to the T score (μX≈ μT), which results in the δ2X value also tending to be close to the δ2T value. It causes the reliability of ρ to approach 1.00. This fact supports the statement that the more test repetitions, the test's reliability tend to increase. It is impossible to do an infinite number of tests, including the online test. But by doing several tests, several repetitions are considered entirely rational and possible to carry out. This study's results prove that repetition of online tests is quite beneficial to increase the accuracy of measurement results. The results also confirm that repetition of three times using the same test, or using the test-retest technique, is proven to increase the online test's reliability. A repeat of the test can bring the measured score closer to the true score, representing the test taker's actual ability. Gamper et al. (2018) state that the test-retest technique is a widely used way to obtain a more consistent test result score. The repetition time interval used is two weeks in the study. Theoretically, the time interval of 2 weeks between the implementation of the 1st, 2nd, and 3rd tests is ideal for using the test-retest technique. The time interval that is too short causing test takers to remember the previous iteration questions. Conversely, a time interval that is too long will allow a change in behavior in the measured aspects. A time interval that is too long can also cause a decrease in the reliability of measurement results due to the increase in information obtained by test takers (Hajjar, 2018; Mardapi, 2012; Widoyoko, 2011). Dias et al. (2019) find that online questionnaires with repetition techniques at 7-11 day intervals produced good reliability. Simões et al. (2018) find that the Pelvic Girdle Questionnaire (PGQ-Brazil) instrument repeatedly in 7-day intervals tends to have good reliability. Roy (2010) also find that the 2-week time interval for the test-retest repetition was ideal. Finally, Marx et al. (2003) argue that different time intervals have relatively no effect on the test-retest reliability. Graph 1 shows that the tendency of increasing reliability will slow down or decrease, in line with the number of repetitions. Although this study only carries out three repetitions, repetitions above three times will likely increase the reliability and a more gentle increase curve. It proves that the repetition of 3 times is relatively sufficient to achieve high reliability. Hart (2017) even finds that the body composition assessment test repeated twice is relatively stable. Roy (2010) believe that the Quebec-French version of the Survey of Pain Attitudes (AD / F-SOFA) test tested twice is relatively stable and reliable. Parraca et al. (2011) also find that the use of the Biodex Balance test in the elderly tended to be reliable at two repetitions. Gravesande et al. (2019) argue that the questionnaire they use is a 4-repetition technique with several variations of the repetition with high-reliability time interval. Ajayi (2013) finds that the reliability calculation method using the test-retest technique on the test on Agricultural Science has a higher degree of reliability than the similar test technique's reliability. The study results also prove that the number of test items affects the reliability of the test results. Table 3 shows that reliability tends to reach adequate numbers on online tests with a higher number of items. Test length is a factor that can increase test reliability (Livingston et al., 2018; Naga et al., 1992). A test that is too short tends to reduce the reliability of the test (Aries, 2011). Sax and Reiter (1980) proves that increasing the grains from 25 to 50 and then to 100 increases the reliability from 0.50 to 0.67, and then to 0.80. Sax concludes that increasing the number of test items increases the test's reliability, although not linearly. Hajjar (2018) proves that reducing test items cannot permanently reduce reliability. If unidimensional items are being deleted, reducing the number of items will reduce reliability. It demonstrates the importance of unidimensional requirements in adding test items to increase test reliability. Increasing the length of the test, or adding test items, will increase the chances of the test being able to cover more test material. It means that the material being measured is also broader. In contrast, shorter tests are generally less able to calculate a sample of behavior (Sax & Reiter, 1980). In measuring learning outcomes, increasing the number of items will lead to more aspects of learning outcomes that can be measured more comprehensively. It will improve the test's performance to describe the actual ability of the object being measured, which means it will increase the test's reliability. Tests that can accommodate a broader range of learning materials will tend to be more reliable. Increasing interrelated items will also reduce the

likelihood of test-takers making guesses. Refraction can occur due to the guess factor in the online test with a multiple-choice model. One way to minimize bias due to guesswork is to create multiple test items measuring the same thing but with different question patterns. An example is the use of positive questions on specific items combined with negative questions on other things. These two items will mutually confirm the correctness and consistency of the test taker's answers. Consistent answers will increase the reliability of the test. Increasing the number of test items also tends to increase test-takers variance (Naga et al., 1992; Sax & Reiter, 1980). If it is related to the Alpha reliability formula from Cronbach (ρa), the total score variance will increase the reliability. It can be seen from the following Alpha Cronbach formula:

$$\rho_a = \frac{k}{k-1}\left(1 - \frac{\Sigma\delta^2_i}{\delta^2_t}\right) \quad \text{..... (Mardapi, 2012; Naga et al., 1992; Sax \& Reiter, 1980)}$$

From the formula above, it can be seen that an increase in the total variance value (δ2t) will cause a decrease in the value of the ratio Σδ2i / δ2t so that the value 1 - (Σδ2i / δ2t). Although the number of item variances (Σδ2) also increases, the increase is lower than the total variance value. On the other hand, the decrease in the fixed multiplier k / (k-1) value will tend to slow down if the number of items k is added. It proves that increasing the length of the test will increase the reliability of the test. Repeated use of online testing, or using test-retest techniques, also has drawbacks. One of them is the carry-over effect factor, which is the tendency for test-takers to respond correctly to the test because they have taken the same test several times (Mardapi, 2012; Sax & Reiter, 1980). Widoyoko (2011) suggests that the test-retest technique should measure learning outcomes in the non-knowledge domain. However, the test-retest method can measure knowledge by changing and randomizing the item numbers on the next test repetition. It is done to reduce the carry-over effect. The importance of using unidimensional items and setting the ideal time interval for the test is also suggested.

Conclusion

Based on the study results using an online test on four subjects, namely the Elementary School Learning Evaluation course, Quantitative Research Methodology, Communication Skills, and Basic Natural Sciences courses, repeated three times, several conclusions are obtained. The results show (1) The online test score used for three repetitions is relatively stable. At the 5% significance level, there is a significant positive correlation between the test scores on the 1st, 2nd, and 3rd repetitions. The correlation coefficient between the score for the 1st repeat test and the 2nd repeat test is in the range of 0.53-0.87. The correlation coefficient between the 1st and 3rd repeat test results is in the field of 0.49-0.77. Simultaneously, the correlation coefficient between the 2nd and 3rd repetition test results is in the range of 0.55-0.84. All correlation coefficients are significant at the 5% level. It proves that the online test result score is relatively stable from one iteration to another. The stability also represents the reliability of the test results. (2) The reliability or internal consistency of online tests tends to show improvement if used repeatedly. Reliability has increased in each course and on average for all classes. In the first repetition, the average reliability of the online tests used is 0.55. In the second repetition, the average reliability is 0.71 and 0.77 on the 3rd iteration. The reliability of the online test reaches a sufficient number on the 3rd iteration. On the 1st iteration, the reliability of the test is in the range of 0.25-0.77. On the 2nd iteration, the reliability is in the range 0.52-0.89. Whereas in the 3rd iteration, the online test's reliability was in the range of 0.67-0.92. (3) Descriptively, the number of test items or the test's length affects the online test's reliability. When using 30 test items, the average reliability is 0.49 and 0.62. When using 40 test items, average reliability of 0.72 is obtained. When using 50 test items, the average test reliability is 0.72. Meanwhile, the number of test-takers does not affect the reliability of online tests. The research implies that it is advisable to repeat the test-retest technique three times to obtain adequate online test results. Using unidimensional online test items and changing the test serial number is also recommended to reduce the carry-over effect.

Acknowledgment

Thank you to the Teacher Training and Education Faculty leaders, as well as the head of the Elementary School Teacher Education study program and the Information Technology Education

study program, Muhammadiyah University of Palangka Raya, for their support in making this research possible. Thanks are also extended to all those who contributed to the process and publication of this research.

Bibliography

Ajayi, B. (2013). A Comparative Analysis of Test Re-Test and Equivalent Reliability Methods. International Journal of Education and Research, 1(6), 1-8. doi: http://www.ijern.com/journal/June-2013/37.pdf

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing Test-Retest Reliability of Psychological Measures. European Psychologist, 22(4), 207-218. doi: 10.1027/1016-9040/a000298

Argaheni, N. (2020). systematic review: The impact of online lectures during the covid-19 pandemic against Indonesian students. . PLACENTUM: Scientific Journal of Health and Its Applications, 8(2), 99-109.

Aries, E. F. (2011). Assessment and Evaluation. Malang: Aditya Media Publishing.

Azwar, S. (1987). Achievement Test, Function and Development of Learning Achievement Measurement, . Yogyakarta: Liberty.

Berchtold, A. (2016). Test–retest: Agreement or reliability? Methodological Innovations, 9, 1-7. doi: 10.1177/2059799116672875

Dias, K., White, J., Metcalfe, C., Kipping, R., Papadaki, A., & Jago, R. (2019). Acceptability, internal consistency and test–retest reliability of scales to assess parental and nursery staff's self-efficacy, motivation and knowledge in relation to pre-school children's nutrition, oral health and physical activity. Public Health Nutrition, 22(6), 967-975. doi: 10.1017/S1368980018004111

Diningrat, S. W. M., Nindya, M. A., & Salwa, S. (2020). EMERGENCY ONLINE TEACHING: EARLY CHILDHOOD EDUCATION LECTURERS'PERCEPTION OF BARRIER AND PEDAGOGICAL COMPETENCY. Journal of Education Horizon, 39(3), 705-719. doi: https://doi.org/10.21831/cp.v39i3.32304

Finch, W. H., & French, B. F. (2018). Educational and psychological measurement: Routledge.

Firman, F., & Rahayu, S. (2020). Online learning in the midst of the covid-19 pandemic. Indonesian Journal of Educational Science (IJES), 2(2), 81-89. doi: https://doi.org/10.31605/ijes.v2i2.659

Gamper, E.-M., Holzner, B., King, M. T., Norman, R., Viney, R., Nerich, V., & Kemmler, G. (2018). Test-Retest Reliability of Discrete Choice Experiment for Valuations of QLU-C10D Health States. Value in Health, 21(8), 958-966. doi: https://doi.org/10.1016/j.jval.2017.11.012

Gravesande, J., Richardson, J., Griffith, L., & Scott, F. (2019). Test-retest reliability, internal consistency, construct validity and factor structure of a falls risk perception questionnaire in older adults with type 2 diabetes mellitus: a prospective cohort study. Archives of Physiotherapy, 9(1), 14. doi: 10.1186/s40945-019-0065-4

Hajjar, S. (2018). Statistical analysis: Internal-consistency reliability and construct validity. International Journal of Quantitative and Qualitative Research Methods, 6(1), 27-38. doi: www.eajournals.org

Hart, P. D. (2017). Test-retest stability of four common body composition assessments in college students. Journal of Physical Fitness, Medicine & Treatment in Sports, 1(2), 1-4. doi: https://www.fitmetrics.org/stability.pdf

Hasanah, U., Fitri, N. L., Supardi, S., & Livana, P. (2020). Depression Among College Students Due to the COVID-19 Pandemic. Journal of Mental Nursing (JKJ): Indonesian National Nurses Association, 8(4), 421-424. doi: https://doi.org/10.26714/jkj.8.4.2020.421-424

Heise, D. R. (1969). Separating Reliability and Stability in Test-Retest Correlation. American Sociological Review, 34(1), 93-101. doi: 10.2307/2092790

Hinton, P., Brownlow, C., McMurray, I., & Cozens, B. (2004). SPSS, Explained. East Sussex, England: Routledge Inc.

Jimenez, L. (2020). Student Assessment During COVID-19. Center for American Progress, 1-7. doi: https://files.eric.ed.gov/fulltext/ED610407.pdf

Kemdikbud. (2020). To prevent the spread of Covid-19 in the education unit, the Ministry of Education and Culture cooperates with the private sector to prepare an online learning solution. from https://www.kemdikbud.go.id/main/blog/2020/03/cegah-sebaran-covid19-di-satuan-pendidikan-kemendikbud-gandeng-swasta-siapkan-solusi-belajar-darin

Livingston, S. A., Carlson, J., Bridgeman, B., Golub-Smith, M., & Stone, E. (2018). Test reliability-basic concepts. Research Memorandum No. RM-18-01). Princeton, NJ: Educational Testing Service. doi: https://www.ets.org/Media/Research/pdf/RM-18-01.pdf

Mardapi, D. (2012). Measurement, assessment, and evaluation of education. Yogyakarta: Nuha Medika.

Marx, R. G., Menezes, A., Horovitz, L., Jones, E. C., & Warren, R. F. (2003). A comparison of two time intervals for test-retest reliability of health status instruments. Journal of Clinical Epidemiology, 56(8), 730-735. doi: https://doi.org/10.1016/S0895-4356(03)00084-2

Naga, N. A.-E., Geiger, K., & Muller, B. (1992). QCD phenomenology of nucleon-nucleon cross sections. Journal of Physics G: Nuclear and Particle Physics, 18(5), 797-805. doi: 10.1088/0954-3899/18/5/009

Parraca, J. A., Olivares Sánchez-Toledo, P. R., Carbonell Baeza, A., Aparicio García-Molina, V. A., Adsuar Sala, J. C., & Gusi Fuertes, N. (2011). Test-Retest reliability of Biodex Balance SD on physically active old people. Journal of Human Sports Exercise, 6(2), 444-451. doi: http://dx.doi.org/10.4100/jhse.2011.62.25

Pietro, G. D., Biagi, F., Costa, P., Karpiński, Z., & Mazza, J. (2020). The likely impact of COVID-19 on education: Reflections based on the existing literature and recent international datasets. JRC Working Papers. Joint Research Centre (Seville site).

Rahiem, M. D. (2020). The emergency remote learning experience of university students in Indonesia amidst the COVID-19 crisis. International Journal of Learning, Teaching and Educational Research, 19(6), 1-26.

Rahim, A. F. A. (2020). Guidelines for online assessment in emergency remote teaching during the COVID-19 pandemic. Education in Medicine Journal, 12(2), 59–68. doi: https://doi.org/10.21315/eimj2020.12.2.6

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2020). Online university teaching during and after the Covid-19 crisis: Refocusing teacher presence and learning activity. Postdigital Science and Education, 2(3), 923-945. doi: https://doi.org/10.1007/s42438-020-00155-y

Roy, D. D. (2010). Cluster analysis for test-retest reliability. International Journal of Psychological Research, 3(1), 131-139. doi: https://doi.org/10.21500/20112084.858

Sax, G., & Reiter, P. B. (1980). Reliability and Validity of Two-Option Multiple-Choice and Comparably Written True-False Items. doi: https://eric.ed.gov/?id=ED236177

Simões, L., Teixeira-Salmela, L. F., Magalhães, L., Stuge, B., Laurentino, G., Wanderley, E., . . . Lemos, A. (2018). Analysis of Test-Retest Reliability, Construct Validity, and Internal Consistency of the Brazilian Version of the Pelvic Girdle Questionnaire. Journal of Manipulative and Physiological Therapeutics, 41(5), 425-433. doi: https://doi.org/10.1016/j.jmpt.2017.10.008

Spencer, F. (2003). The School-Years Screening Test for the Evaluation of Mental Status Test-Reliability and Cognitive Functioning Stability. Paper presented at the Proceedings of the 38th APS Annual Conference-Development through Diversity.

Sukandi, P., Kautsar, M. E., Zidane, M. Y., Permana, T., & Alfaruqi, S. (2021). Analysis of Student Social Interactions During the Covid-19 Pandemic (Case Study of D3 Management Student of Widyatama University). Psychology and Education Journal, 58(3), 377-381. doi: https://doi.org/10.17762/pae.v58i3.2731

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research;How to test the validation of a questionnaire/survey in a research. International Journal of Academic Research in Management (IJARM), 5(3), 28-36. doi: https://dx.doi.org/10.2139/ssrn.3205040

Wells, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. Testing & Evaluation Services., 1-7. doi: https://testing.wisc.edu/Reliability.pdf

Widoyoko, E. P. (2011). The evaluation of the learning programme a practical guide for educators and prospective educators. Yogyakarta: Pustaka Pelajar.