



## SAMPLE SIZE DETERMINATION IN TEST-RETEST AND CRONBACH ALPHA RELIABILITY ESTIMATES

Imasuen Kennedy

University of Benin, Institute of Education, Nigeria

kennedy.imasuen@uniben.edu, +234 8109670163

### Cite this article:

Imasuen Kennedy (2022), Sample Size Determination in Test-Retest and Cronbach Alpha Reliability Estimates. British Journal of Contemporary Education 2(1), 17-29. DOI: 10.52589/BJCE-FY266HK9

### Manuscript History

Received: 29 Dec 2021

Accepted: 25 Jan 2021

Published: 3 Feb 2022

### Copyright © 2022 The Author(s).

This is an Open Access article distributed under the terms of Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0), which permits anyone to share, use, reproduce and redistribute in any medium, provided the original author and source are credited.

**ABSTRACT:** *The estimation of reliability in any research is a very important thing. For us to achieve the goal of the research, we are usually faced with the issue of when the measurements are repeated, are we sure we will get the same result? Reliability is the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials. If a measure is perfectly reliable, there is no error in measurement, that is, everything we observe is the true score. However, it is the amount/degree of error that indicates how reliable, a measurement is. The issue of sample size determination has been a major problem for researchers and psychometricians in reliability studies. Existing approaches to determining sample size for psychometric studies have been varied and are not straightforward. This has made the psychometric literature contain a wide range of articles that propose a variety of sample sizes. This paper investigated sample sizes in test-retest and Cronbach alpha reliability estimates. The study was specifically concerned with identifying and analyzing differences in test-retest and Cronbach alpha reliability estimate of an instrument using various sample sizes of 20,30,40,50,100,150,200,300, and 400. Four hundred and eight (408) senior secondary school students from thirty-eight (38) public senior secondary schools in Benin metropolis part took in the study. The Open Hemisphere Brain Dominance Scale, by Eric Jorgenson was used for data collection. Data were analyzed using Pearson Product Moment Correlation Coefficient (r) and Cronbach alpha. The findings revealed that the sample sizes of 20 and 30 were not reliable, but the reliability of the instrument became stronger when the sample size was at least 100. The interval estimate (Fisher's confidence interval) gave a better reliability estimate than the point estimate for all samples. Based on the findings, it was, therefore, recommended that for a high-reliability estimate, at least one hundred (100) subjects should be used. Observed or field-tested values should always be used in the estimation of the reliability of any measuring instrument, and reliability should not be reported as a point estimate, but as an interval.*

**KEYWORDS:** Reliability, Sample size, Test-retest, Cronbach Alpha



## INTRODUCTION

The estimation of reliability and validity in any research is very important. For us to achieve the goal of the research, we are usually faced with two issues; the first is how do we ascertain that we are indeed measuring what we want to measure?”, and “if we repeat the measurement, are we sure we will get the same result?” The first question is related to the issues of validity and the second to reliability. These two concepts are referred to as psychometric properties.

The term reliability in psychological research refers to the consistency of a research study or measuring test (McLeod, 2007). If findings from research can be replicated consistently, they are reliable. Most times obtaining the same results may not be feasible as participants and situations vary. However, if a strong positive correlation exists between the results of the same test, this indicates reliability (Balkin, 2017).

Many definitions abound in the literature of psychometrics of reliability. According to Wilkinson and Robertson (2006) reliability with respect to research means "repeatability" or "consistency". Reliability can also be defined as the degree to which an assessment tool produces stable and consistent results (Meyer, 2010). On his part Mellenbergh, (2011) opined that reliability is the consistency of a test or the degree to which the test gives consistent results. It is also seen as a measure of a test's precision. Reliability is the extent to which an experiment, test, or any measuring procedure yields the same result on repeated trials.

According to National Council on Measurement in Education (NCME; 1999), reliability in statistics and psychometrics is the overall consistency of a measure. A measure is said to have high reliability if it produces similar results under consistent conditions. It is the characteristic of a set of test scores that relates to the amount of random error from the measurement process that might be embedded in the scores. Highly reliable scores are accurate, reproducible, and consistent from one testing occasion to another. That is, if the testing process were repeated with a group of test-takers, essentially the same results would be obtained.

According to the standards written by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME), (2014), reliability refers to the consistency of measurements when a testing process is repeated for an individual or group of individuals.

Reliability is the extent to which a questionnaire, test, observation or any measurement procedure produces the same results on repeated trials (Bolarinwa, 2015). In short, it is the stability or consistency of scores over time or across raters (Miller, 2015). It is worthy to note that lack of reliability may arise from divergences between observers or instruments of measurement or instability of the attribute being measured (Last, 2015). Nunnally, (cited in Bardhoshi, et al 2016) opined that measurements are reliable to the extent that they are repeatable and that any random influence that tends to make measurements different from occasion to occasion or circumstance to circumstance is a source of measurement error.

According to Kline (2000), reliability, as it applies to test, has two distinct meanings. One refers to stability over time, the second to internal consistency. Reliability is the degree to which a test consistently measures whatever it measures. Reliability is an indicator of consistency, that is, an indicator of how stable a test score or data is across applications or time. A measure should produce similar or the same results consistently if it measures the same “thing.”



(Sawilowsky, 2000). A measure can be reliable without being valid but a measure cannot be valid without being reliable (Erford, 2013).

The correlation coefficient plays an important role in the determination of the degree of reliability. A correlation coefficient of + 1.0 is regarded as a perfect positive relationship, - 1.0 as a perfect negative relationship and that of 0.0 indicates no relationship. The nearer a correlation is to +1.0, the more reliable the results. If a measure is perfectly reliable, there is no error in measurement, that is, everything we observe is a true score. Therefore, for a perfectly reliable measure, the reliability = 1. Now, if we have a perfectly unreliable measure, there is no true score, that is, the measure is entirely in error. In this case, the reliability = 0. The value of a reliability estimate tells us the proportion of variability in the measure attributable to the true score. A reliability of 0.5 means that about half of the variance of the observed score is attributable to truth and half is attributable to error. According to American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) (2014) a reliability of 0.8 means the variability is about 80% true ability and 20% error. All measurement procedures involve error. However, it is the amount/degree of error that indicates how reliable measurement is. When the amount of error is low, the reliability of the measurement is high. Conversely, when the amount of error is large, the reliability of the measurement is low, (Elford, 2013; Meyer, 2010).

It is fundamental to note that reliability refers to the result and not the test itself. The samples from which the reliability coefficient are derived must be representative of the population for whom the test is designed and sufficiently large to be statistically reliable (Leann, & Ken, 2012). According to Kline (2000), reliability of 0.7 is a minimum for a good test. This is simply because the standard error of measurement (which is the estimated standard deviation of scores) of scores increases as the reliability decreases.

In general, there are four broad types of reliability: test-retest reliability, parallel forms reliability, internal consistency of reliability, and inter-rater reliability (Kaplan & Saccuzzo, 2005). In this study, we shall examine stability (test-retest) and internal consistency (Cronbach alpha).

### **Test-retest Reliability (or Stability)**

Test-retest reliability (also called Stability) answers the question, “will the scores be stable over time?” Test-retest reliability refers to the temporal stability of a test from one measurement session to another. The procedure is to administer the test to a group of respondents and then administer the same test to the same respondents at a later date. The correlation between scores on the identical tests given at different times operationally defines its test-retest reliability. Two assumptions underlie the use of the test-retest procedure; (Wells, 2003)

- The first required assumption is that the characteristic that is measured does not change over the time period called 'testing effect' (Engel & Schutt, 2013)
- The second assumption is that the time period is long enough yet short in time that the respondents' memories of taking the test, the first time does not influence their scores at the second time and subsequent test administrations called 'memory effect'.

The estimate of test-retest reliability is also known as the coefficient of stability (Cohen et al, 1996). Test-retest correlation provides an indication of stability over time (Wong, Ong & Kuek,



2012, Pedisic *et al*, 2014; Deniz, & Alsaffar, 2013). In other words, the scores are consistent from the first administration to the second administration. In using this form of reliability, one needs to be careful with questionnaires or scales that measure variables that are likely to change over a short period of time, such as energy, happiness and anxiety because of the maturation effect (Drost, 2011). For well-developed standardized achievement tests administered reasonably close together, test-retest reliability estimates tend to range between 0.70 and 0.90 (Popham, 2000)

Despite its appeal, the test-retest reliability technique has several limitations (Rosenthal & Rosnow, 1991). For instance, when the interval between the first and second test is too short, respondents might remember what was on the first test and their answers on the second test could be affected by memory. Alternatively, when the interval between the two tests is too long, maturation happens. Kaplan and Saccuzzo (2005) noted that test-retest reliability estimates evaluate the reliability of instrument scores when an instrument is given at multiple and subsequent points in time. Joppe, (2000) detects a problem with the test-retest method which can make the instrument, to a certain degree, unreliable. She explains that the test-retest method may sensitize the respondent to the subject matter, and hence influence the responses given. Similarly, Crocker and Algina (1986) noted that when a respondent answers a set of test items, the score obtained represents only a limited sample of behaviour.

### **Internal Consistency**

Internal consistency reliability answers the question, “How well does each item measure the content or construct under consideration?” The appeal of an internal consistency index of reliability is that it is estimated after only one test administration and, therefore, avoids the problems associated with testing over multiple time periods. (Wong, Ong, & Kuek, 2012). The internal consistency reliability estimate refers to the inter-correlations between items on the same instrument (Kaplan & Saccuzzo, 2005). Cronbach’s coefficient alpha is one of the most frequently used ways of estimating internal consistency of reliability (Dimitrov, 2002). The  $\alpha$  coefficient is the most widely used procedure for estimating reliability in applied research. As stated by Sijtsma (2009), its popularity is such that Cronbach (1951) has been cited as a reference more frequently than the article on the discovery of the DNA double helix. Nevertheless, its limitations are well known (Yang & Green, 2011), some of the most important being the assumptions of uncorrelated errors, tau-equivalence and normality

### **Sample size determination in reliability**

The issue of sample size determination has been a major problem for researchers and psychometricians in the reliability study. Existing approaches to determining sample size for psychometric studies have been varied and are not straightforward. This has made the psychometric literature to contain a wide range of articles that propose a variety of sample sizes (Donner & Eliasziw 1987; Eliasziw *et al*, 1994; Cocchetti, (1999); Charter, (1999); Mendoza, Stafford, & Stauffer, (2000); Bonett, 2002). These studies are classified into two broad categories: those based on authors’ experiences and those on statistical theory.

In the studies based on judgments from authors’ experiences (DeVellis, 1991; Rea, & Parker, 1992; Ferguson, & Cox, 1993), the sample size recommendations vary widely. Other authors advocated and suggested that samples should exceed 300 (Ware, *et al*, (1997), whereas some posited that much smaller samples as little as 30 subjects (Rea, & Parker, 1992; Bonett &



Wright, 2014) may suffice. The second category of sample size recommendations includes those studies grounded in statistical theory (Feldt, et al, 1987; Donner & Eliasziw, 1987; Eliasziw, et al, 1994; Bonett, (2002). These differ in approaches for reliability testing (Charter, 1999; Mendoza et al, 2000) and recommendations ranging from  $n = 25$  (Cocchetti, 1999) to 400 for reliability testing (Charter, 1999).

Kline, (2000) advised that researchers should use at least 100 participants per item on our scale if the reliability estimate is to be meaningful. A lot of surprising differences of opinion on sample size determination abound in the literature. Some authors are suggesting that samples as small as thirty (30) (Bonett, & Wright, 2014), can measure the reliability, so long as the scale items have strong inter-correlation. Toe-ing the same line, Nunnally & Bernstein (1994) averred that the minimum criteria for reliability coefficients for Cronbach's Alpha are 0.80; 0.30 for item-total correlations, 0.30 for item-item correlations, and 0.80 for intra-class correlation coefficients. Kline (1986) suggested a minimum sample size of 300, as did Nunnally & Bernstein (1994). Segall (1994) called a sample size of 300 "small". Charter (1999) stated that a minimum sample size of 400 was needed for a sufficiently precise estimate of the population coefficient alpha. Charter (2003) opined that with low sample sizes alpha coefficients can be unstable. Walker and Zhang (2004) suggested a minimum sample size of 125 to 150 for calculating reliability, with at least as many people in the sample as items on the test. However, the minimum sample size for the sample coefficient alpha has been frequently debated due to the difficulty of data collection in psychometric research. Although the determination of the sample size needed for reliability studies is somewhat subjective, a minimum of 400 subjects is recommended.

In reliability studies, various sample sizes are used by different authors and researchers. Furthermore, there is no uniformity in the sample sizes been used. Sample size plays an important role in the estimation of the reliability level of the measurement scale.

Correlations, along with most other statistical indices, have standard errors, indicating how trustworthy the results are. However, it can be said that the larger the number of subjects the smaller the standard error of the statistics (Erford, 2013). This means that it is essential that the reliability estimates are derived from a sample sufficiently large to minimize this statistical error (AERA, APA, & NCME, 2014). In reliability testing, determining the right sample size is oftentimes critical (Erford, 2013; Meyer, 2010). If the sample size used is too small, not much information can be obtained from the test, thereby limiting one's ability to draw meaningful conclusions. On the other hand, if it is too large, information obtained through the test may be beyond what is needed (AERA, APA, & NCME, 2014). Thus, incurring unnecessary costs. But most times, the test developers do not have the luxury to request how many samples are needed but has to create a test plan based on the budget or resource constraints that are in place for the project.

### **Statement of the Problem**

There is a surprising difference of opinion in literature as regards the adequate sample size for establishing the reliability of research instruments. For example, Kline (2000) noted that the standard advice is to use at least 100 participants per item on our scale if the reliability estimate is to be meaningful. On the other hand, Bonnet and Wright (2014) asserted that samples must be as small as thirty (30) to establish reliability so long as the scale items have strong inter-correlation. More so, many researchers use different sample sizes for establishing reliability





estimates when carrying out research studies. Some use 20, 30, 40, 50 or 100 samples as the case may be. But no scientific research has been carried out to justify the usage of these sample sizes. Also, some researchers use different methods to establish the different types of reliability. For example, some use test-retest for questionnaire instrument as against the popular Cronbach alpha (Vacha-Haase & Thompson 2010).

Although the topic reliability has gained much attention in the literature, investigations into sample size requirements remain scarce. It is, therefore, imperative to examine the test-retest and Cronbach alpha (the most used reliability estimates) of an instrument using various sample sizes.

### **Research Questions**

The following research questions were raised to guide the study.

1. Is there a difference in the test-retest reliability estimate of an instrument using various sample sizes of 20, 30, 40, 50, 100, 150, 200, 300, 400?
2. Is there a difference in the Cronbach alpha reliability estimate of an instrument using various sample sizes of 20, 30, 40, 50, 100, 150, 200, 300, 400?

### **Relevance of the Study**

The findings of the study will help psychometricians, educators and researchers to be aware of the minimum sample size in carrying out reliability studies. This will put to an end the problem of choosing the right sample size for acceptable reliability. It will be an eye-opener to psychometricians and researchers on the method and sample size to use when conducting a reliability study. In the same vein, the findings will help psychometricians and researchers to estimate the proportion of variability in their measurement which is attributable to the true score. That is, it will help them to determine the amount /degree of error which indicates how reliable a measure is. When the amount of error is low, the reliability of the measurement is high and conversely, when the amount of error is large, the reliability of the measure is low.

This study will also be beneficial to researchers and other stakeholders who may be having problems with choosing the appropriate methods of estimating reliability estimates. And this study will help all researchers and other stakeholders to report accurately reliability estimates in any manuscripts (test manuals, conference papers and articles)

### **Methods**

The survey research design was adopted for the study. The population of this study comprised of all the students in public Senior Secondary School in Benin metropolis in Edo state. A total of seventy-five (75) senior secondary schools with a total number of 40,815 students is in the Benin metropolis. The breakdown is as follows: Egor Local government area 12 schools with 8,207 students; Oredo local government area have 13 thirteen senior secondary schools with 12,154 students; Ikpoba Okha local government area have 27 senior secondary schools with 15,456 students and Ovia North East with 23 senior secondary school and 4998 students. The statistics of schools and students were collected from the Ministry of Education, Benin City. A sample size of 408 students from senior secondary schools was selected from thirty-eight (38) senior secondary schools in Benin metropolis. The multistage sampling technique which



involves various sampling stages was used for selecting the samples. The instrument for data collection was the Open Hemisphere Brain Dominance Scale 1.0 (OHBDS), a personality scale designed by Eric Jorgenson (2015). This was adapted by the researcher. It consists of two sections. Section A was used to elicit information from the student biodata, which includes their sex, and class. Section B consists of a twenty (20) items inventory designed to measure the hypothesized left-brain versus right-brain preference among students with a 4 - point Likert scale. The items are under the options of response: SA = Strongly Agree, A = Agree, D = Disagree, SD = Strongly Disagree. SD will be scored 1 point, D was scored 2 points, A was scored 3 points and SA scored 4 points. The instrument has been validated by Eric Jorgenson but was also validated by experts in Measurement and Evaluation, University of Benin, Benin City. The reliability of the instrument was part of the issues raised in the study.

The reliability coefficient was estimated using the Pearson Product Moment Correlation Coefficient ( $r$ ) for the instrument that was subjected to test re-test, and Cronbach alpha  $\alpha$ , for the instrument that was administered once. The Fisher's 95% confidence interval was used to determine which of the sample sizes give a stable result. The width of the interval for the various sample sizes was determined. The sample size(s) with a shorter interval was adjudged as the most stable and consistent

## RESULTS

**Table 1: Fisher 95% Confidence Interval of Test Retest Reliability Estimates**

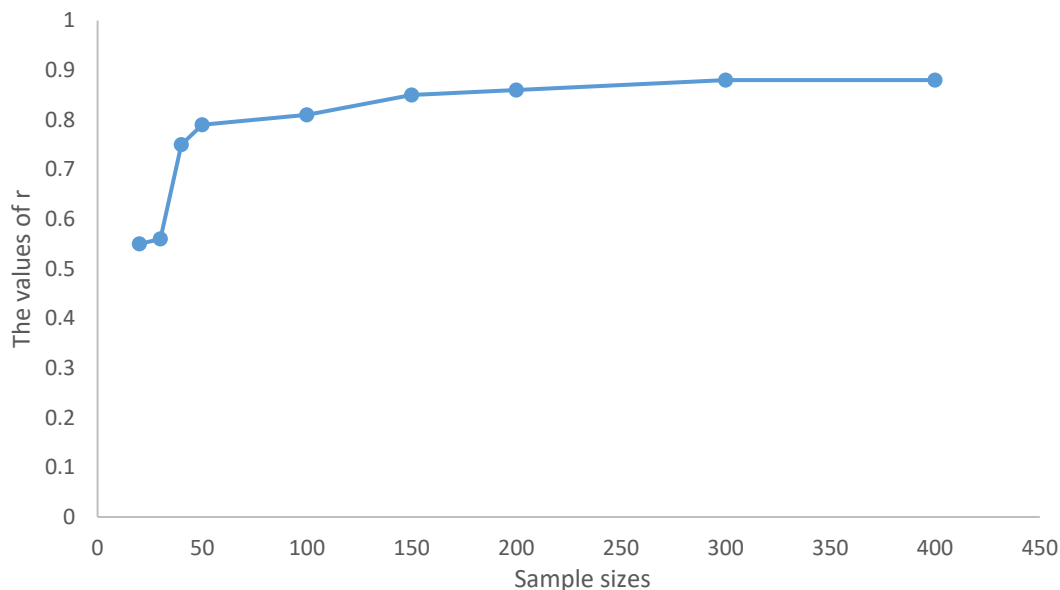
Sample size	$r$	$Z_r$	$\sigma_z$	$\sigma_z(1.96)$	$Z_{rLB}$	$Z_{rUB}$	$\rho_{LB}$	$\rho_{UB}$	Width
20	0.55	0.618	0.243	0.475	0.143	1.093	0.142	0.798	0.66
30	0.56	0.633	0.192	0.376	0.257	1.009	0.251	0.765	0.51
40	0.75	0.973	0.164	0.321	0.652	1.294	0.573	0.860	0.29
50	0.79	1.071	0.146	0.286	0.785	1.357	0.656	0.876	0.22
100	0.81	1.127	0.102	0.199	0.928	1.326	0.730	0.868	0.14
150	0.85	1.256	0.082	0.161	1.095	1.417	0.799	0.889	0.09
200	0.86	1.293	0.071	0.139	1.154	1.432	0.819	0.892	0.07
300	0.88	1.376	0.058	0.114	1.262	1.490	0.852	0.903	0.05
400	0.88	1.376	0.050	0.098	1.278	1.474	0.856	0.900	0.04

Key:  $r$  = Pearson  $r$ ;  $Z_r$  = Fisher  $Z$ ;  $\sigma_z$  = Standard Error of Fisher  $Z$ ;  $Z_{rLB}$  = Lower bound of Fisher  $Z$ ;  $Z_{rUB}$  = Upper Bound of Fisher  $Z$ ;  $\rho_{LB}$  = Lower bound of Pearson  $r$ ;  $\rho_{UB}$  = Upper Bound of Pearson  $r$

The result in Table 1 showed the Fisher 95% confidence interval of test retest reliability estimates for an instrument using various sample sizes of 20,30, 40,50,100,150,200,300, and 400. It further shows that with a sample size of 20, the  $r$  value was 0.55, with a 95% confidence interval of (0.14, 0.80) and a width of 0.66. When the sample was increased to 30 the  $r$  value became 0.56 with a 95% confidence interval of (0.25, 0.77) and a width of 0.52. A sample size of 40 gave an  $r$  value of 0.75 with a 95% confidence interval of (0.57, 0.86) and a width of 0.29. A sample size of 50 gave an  $r$  value of 0.79 with a 95% confidence interval of (0.66,



0.88) and a width of 0.22. When the size became 100, the value of  $r$  became 0.81 with a 95% confidence interval of (0.73, 0.87) and a width of 0.14. A sample size of 150 gave an  $r$  value of 0.85 with a 95% confidence interval of (0.80, 0.89) and a width of 0.09. The sample size of 200 gave an  $r$  value of 0.86 with a 95% confidence interval of (0.82, 0.89) and a width of 0.07. 300 samples gave an  $r$  value of 0.88 with a 95% confidence interval of (0.85, 0.90) and a width of 0.05. A sample size of 400 gave an  $r$  value of 0.88 with a 95% confidence interval of (0.86, 0.90) and a width of 0.04. This is presented in figure 1



**Figure 1: Fisher 95% Confidence Interval of Test – Retest Reliability Estimates**

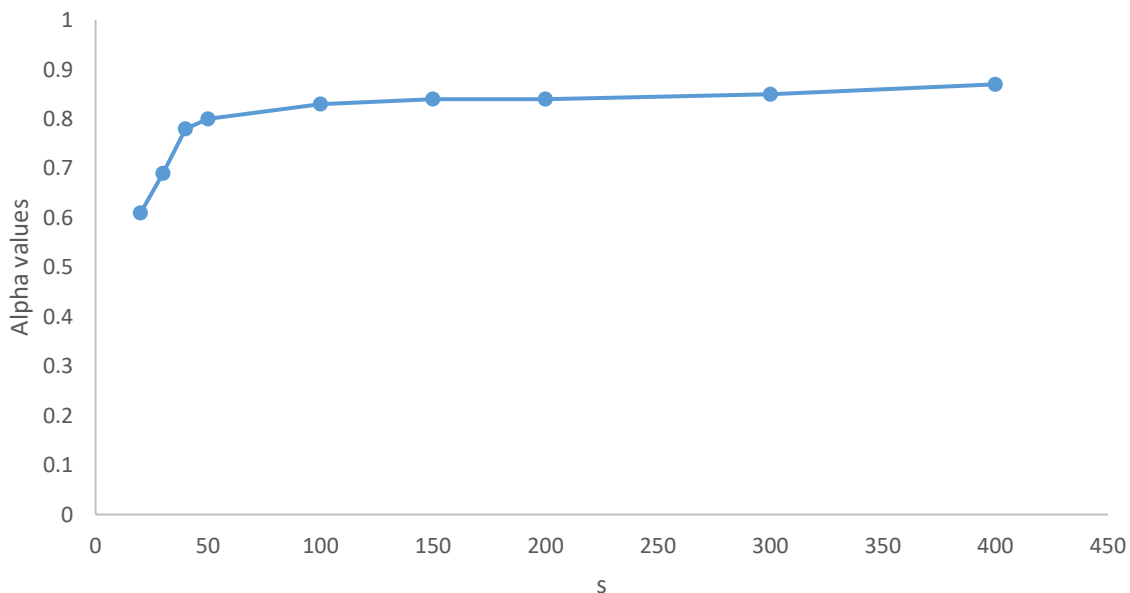
**Table 2: Fisher 95% Confidence Interval of Cronbach Alpha Reliability Estimates**

Sample sizes	$\alpha$	$Z_\alpha$	$\sigma_z$	$Z_\alpha(1.96)$	$Z_\alpha LB$	$Z_\alpha UB$	$\rho LB$	$\rho UB$	Width
20	0.61	0.709	0.243	0.475	0.234	1.184	0.230	0.829	0.60
30	0.69	0.848	0.192	0.376	0.472	1.224	0.440	0.841	0.40
40	0.78	1.045	0.164	0.321	0.724	1.366	0.619	0.873	0.26
50	0.80	1.099	0.146	0.286	0.813	1.385	0.675	0.885	0.21
100	0.83	1.188	0.102	0.199	0.989	1.387	0.757	0.883	0.10
150	0.84	1.221	0.082	0.161	1.060	1.382	0.786	0.880	0.09
200	0.84	1.221	0.071	0.139	1.082	1.360	0.794	0.876	0.08
300	0.85	1.256	0.058	0.114	1.142	1.370	0.815	0.879	0.06
400	0.87	1.333	0.050	0.098	1.235	1.431	0.844	0.892	0.05

Key:  $\alpha$  = Cronbach alpha;  $Z_\alpha$  = Fisher Z;  $\sigma_z$  = Standard Error of Fisher Z;  $Z_\alpha LB$  = Lower bound of Fisher Z;  $Z_\alpha UB$  = Upper Bound of Fisher Z;  $\rho LB$  = Lower bound of Pearson r;  $\rho UB$  = Upper Bound of Pearson r



The result in Table 2 showed the Fisher 95% confidence interval of Cronbach alpha reliability estimates of an instrument using various sample sizes of 20, 30, 40, 50, 100, 150, 200, 300, and 400. It further shows that with a sample size of 20, the  $\alpha$  value was 0.61, with a 95% confidence interval of (0.23, 0.83) and a width of 0.60. When the sample was increased to 30 the  $\alpha$  value became 0.69 with a 95% confidence interval of (0.44, 0.84) and a width of 0.40. A sample size of 40 gave an  $\alpha$  value of 0.78 with a 95% confidence interval of (0.62, 0.87) and a width of 0.26. A sample size of 50 gave an  $\alpha$  value of 0.80 with a 95% confidence interval of (0.68, 0.89) and a width of 0.21. When the size became 100, the value of  $\alpha$  became 0.83 with a 95% confidence interval of (0.77, 0.89). A sample size of 150 gave an  $\alpha$  value of 0.84 with a 95% confidence interval of (0.79, 0.88) and a width of 0.09. The sample size of 200 gave an  $\alpha$  value of 0.84 with a 95% confidence interval of (0.80, 0.88) and a width of 0.08. 300 samples gave an  $\alpha$  value of 0.85 with a 95% confidence interval of (0.82, 0.88) and a width of 0.06. A sample size of 400 gave an  $\alpha$  value of 0.87 with a 95% confidence interval of (0.84, 0.89) and a width of 0.05. This is presented in figure 2



**Figure 2: Fisher 95% Confidence Interval of Cronbach Alpha Reliability Estimates**

## DISCUSSION OF FINDINGS

The study revealed that the sample sizes of 20 and 30 using the test-retest statistics were not reliable. The sample size of 40 and 50, though reliable, the lower bound was outside the acceptable reliability of 0.70 for a test-retest (Kline 2000). The reliability of the instrument became stronger when the sample size was at least 100. This finding is in line with Leann, & Ken, (2012) who affirmed that the samples from which the reliability coefficient are derived must sufficiently be large to be statistically reliable. The finding is also in collaboration with the study of Kline (2000) who noted that the standard advice is to use at least 100 participants per item on our scale if the reliability estimate is to be meaningful. In the same vein, the finding is supported by Ware et al (1997) who asserted that samples should exceed 300. But the finding



disagreed with Bonnet & Wright (2014) who asserted that samples must be as small as thirty (30) to establish reliability so long as the scale items have strong inter-correlation and Rea, & Parker, (1992) who posited that smaller samples as little as 30 subjects may suffice for test-retest reliability.

The study also revealed that the sample sizes of 20 and 30 using the Cronbach alpha statistics were not reliable. The sample size of 40 and 50, though reliable, the lower bound was outside the 0.80 acceptable reliability coefficients for Cronbach's Alpha (Nunnally & Bernstein (1994). The reliability of the instrument became stronger when the sample size was at least 100. This finding is in line with AERA, APA, & NCME, (2014) and Erford, (2013) who stated that the larger the number of subjects the smaller the standard error of the statistic which means that it is essential that the reliability estimates are derived from a sample sufficiently large to minimize this statistical error. The finding is also in collaboration with the study of Kline (1986) who suggested a minimum sample size of 300, as did Nunnally & Bernstein (1994). Segall (1994) called a sample size of 300 "small". Charter (1999) stated that a minimum sample size of 400 was needed for a sufficiently precise estimate of the population coefficient alpha. Charter (2003) also noted that with low sample sizes alpha coefficients can be unstable. Walker and Zhang (2004) suggested a minimum sample size of 125 to 150 for calculating reliability, with at least as many people in the sample as items on the test. Charter, (1999) suggested a sample size of 400 for reliability testing. But the finding disagreed with Feldt et al, (1987), Donner & Eliasziw (1987), Eliasziw et al, (1994), Bonett, (2002), Charter, (1999), Mendoza et al, (2000) and Cocchetti, (1999) who recommended a sample size ranging from  $n = 25$

The difference in the finding of this study could be as a result of using observed values from the field. Most of the findings in the literature were either from personal experience or statistical theorem. Unfortunately, much of the empirical evidence comes from simulated data. So their recommendations are incomplete because simulated data have important limitations as compared to observed data. They are based on preselected statistical or computer models that can only approximate observed data, have artificially controllable parameters, and are often generated to reflect randomly distributed samples. These limit the inferences that can be drawn from analyzing simulated data and necessitate the collection of observed data to ensure their credibility.

Another revelation from the study is that both the test-retest and Cronbach reliability estimates started converging from the sample size of 100 (see figures 1 and 2). This, therefore, implies that for an acceptable reliability study, at least one hundred subjects should be used.

The result of the study also revealed that the interval estimate gave a better reliability estimate than the point estimate for all the samples. For example, for the test-retest, a sample of 40 gave a reliability index of 0.75 as a point estimate, but the interval estimate gave a reliability estimate of (0.573, 0.860). The lower bound was outside the acceptable reliability index of  $\geq 0.70$ . This collaborates with the study of AERA, APA, & NCME, (2014), who advocated reporting reliability estimates as interval estimates against the point estimate previously used.



## CONCLUSION

Based on the finding of this study, the following conclusions emerged. The result demonstrated that a number of differences exist in the sample size determination of a reliability study. The usage of sample sizes of twenty (20) and thirty (30) was not justified. This could be attributed to the fact that other studies that suggested a minimum of 20 and 30 subjects used simulated data as against observed data used in this study.

The larger the number of subjects the smaller the standard error of the statistic. To minimize this statistical error, the reliability estimates must be derived from a sufficiently large sample. The findings of the study have shown that the usage of sample sizes of 20 and 30 for reliability studies is not justifiable. It has also shown that for an acceptable reliability study, the sample size should be at least one hundred (100).

## RECOMMENDATIONS

The reliability of any measuring instrument is a task frequently encountered in research. Sample size determination plays a very important role in the estimation of reliability. The higher the sample, the higher the reliability and the lower the error inherent in the instrument. Based on this, the following recommendations were made.

1. Observed or field-tested values should always be used in the estimation of the reliability of any measuring instrument.
2. For a high-reliability estimate, at least one hundred (100) subjects should be used.
3. Reliability should not be reported as a point estimate but as an interval estimate.

## REFERENCES

- American Educational Research Association (AERA). (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, 35(6), 33–40.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- Balkin, R. S. (2017). Evaluating evidence regarding relationships with criteria. *Measurement and Evaluation in Counseling and Development*, 50, 264–269.
- Bardhoshi, G., Erford, B. T., Duncan, K., Dummett, B., Falco, M., Deferio, K., & Kraft, J. (2016). Choosing assessment instruments for posttraumatic stress disorder screening and outcome research. *Journal of Counseling & Development*, 94, 184–194
- Bolarinwa, O.A (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Niger Postgraduate Medical Journal* 22:195-201
- Bonett, D. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational Behavioural Statistics* 27:335–340
- Bonett, D. G., and Wright, T. A (2014). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size. *Journal of Organizational Behaviour*. 36(1)



- Charter, R. A. (1999). Sample Size Requirements for Precise Estimates of Reliability, Generalizability, and Validity Coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21, 559-566.
- Charter, R.A. (2003). Study Samples Are Too Small to Produce Sufficiently Precise Reliability Coefficients. *The Journal of General Psychology*, 130, 117-129.
- Cocchetti, D. (1999). Sample size requirements for increasing the precision of reliability estimates: problems and proposed solutions. *Journal of Clinical Experimental Neuropsychology* 21:567-570
- Cohen D., Nisbett, R.E., Bowdle, B.F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honour: an “experimental ethnography.” *Journal of Personality and Social Psychology*. 70, 945-60.
- Crocker, L., & Algina, J. (1986) *Introduction to Classical and Modern Test Theory*, Harcourt Brace Jovanovich College Publishers: Philadelphia.
- Deniz, M.S, Alsaffar, A.A (2013). Assessing the validity and reliability of a questionnaire on dietary fibre-related knowledge in a Turkish student population. *Journal of Health Population and Nutrition*, 31:497-503.
- Devellis, R.F. (1991) *Scale Development: Theory and Applications*, Applied Social Research Methods Series 26, Sage: Newbury Park.
- Dimitrov, D.M. (2002). *Error variance of Rasch measurement with logistic ability distributions*. Paper presented at the meeting of the American Educational Research Association. New Orleans, Louisiana.
- Donner, A., & Eliasziw, M. (1987). Sample size requirements for reliability studies. *Journal of Statistical Medicine* 6:441-448
- Drost, E.A (2011). Validity and reliability in social science research. *Educational Research Perspective*; 38:105-23.
- Eliasziw, M., Young, S., Woodbury, M., & Fryday-Field, K. (1994). Statistical methodology for the concurrent assessment of interrater and interrater reliability: using goniometric measurements as an example. *Journal of Physical Therapy* 74:777-788
- Engel, R.J, & Schutt, R.K (2013). *Measurement. The Practice of Research in Social Work*. 3<sup>rd</sup> edition, Sage Publication Inc. (Online); 2013. p. 97-104. Available from: [https://www.us.sagepub.com/sites/default/files/upm-binaries/45955\\_chapter\\_4.pdf](https://www.us.sagepub.com/sites/default/files/upm-binaries/45955_chapter_4.pdf). [Last accessed on 2015 Oct 10].
- Erford, B. T. (2013). *Assessment for counsellors* (2nd ed.). Belmont, CA: Cengage Wadsworth.
- Erford, B. T., Johnson, E., & Bardhoshi, G. (2016). Meta-analysis of the English version of the Beck Depression Inventory-Second edition. *Measurement and Evaluation in Counseling and Development*, 49, 3-33.
- Feldt, L. Woodruff, D., & Sailh, F. (1987). Statistical inference for coefficient alpha. *Journal of Applied Psychological Measure* 11:93-103
- Ferguson, E., & Cox, T. (1993). Exploratory factor analysis: a user’s guide. *International Journal of Selection and Assessment* 1:84-94.
- Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: AERA.
- Joppe, M. (2000). *The Research Process*. Retrieved February 25, 1998, from <http://www.ryerson.ca/~mjoppe/rp.htm>
- Kaplan, R. M. & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues* (6th Ed.). Belmont, CA: Thomson Wadsworth.
- Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. New York: Methune & Company.



- Kline, R. B. (2000). *Beyond significance testing: Reforming data analysis methods in behavioural research*. Washington, DC: *American Psychological Association*.
- Last, J.M. (2015). *A Dictionary of Epidemiology*. 4<sup>th</sup> edition. New York: Oxford University Press.
- Leann, J. T., and Ken, J (2012). Sample size planning for composite reliability coefficients: accuracy in parameter arrow estimation via narrow confidence intervals. *The British Journal of mathematics and statistical psychology*. 65: 371 - 401
- McLeod, S. A. (2007). *What is Reliability?*. Retrieved on 27th June 2017 from [www.simplypsychology.org/reliability.html](http://www.simplypsychology.org/reliability.html).
- Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics*. The Hauge, Netherlands: Eleven International.
- Mendoza, J., Stafford, K., & Stauffer, J (2000). Large-sample confidence intervals for validity and reliability coefficients. *Journal of Psychological Methods* 5:356–369
- Meyer, P. (2010). *Reliability: Understanding statistics measurement*. New York, NY: Oxford University Press.
- Miller, M.J (2015). *Graduate Research Methods*. Available from: <="" a="">. [Last accessed on 2015 Oct 10].
- National Council on Measurement and Evaluation in Education (1999)
- Nunnally, J.C., & Bernstein, I.H. (1994) *Psychometric Theory*, (3<sup>rd</sup> ed), McGraw-Hill: New York.
- Pedisic, Z, Bennie, J.A, Timperio, A.F, Crawford, D.A, Dunstan, D.W, & Bauman, A.E, (2014). Workplace sitting breaks questionnaire (SITBRQ): An assessment of concurrent validity and test-retest reliability. *BMC Public Health* 14:1249
- Popham, W. J (2000). *Modern educational measurement: Practical guidelines for educational leaders*. 3<sup>rd</sup> edition Needham, MA: Allyn and Bacon
- Rea, L., & Parker, R. (1992). *Designing and conducting survey research: a comprehensive guide*. Jossey-Bass, San Francisco
- Rosenthal, R., and Rosnow, R. L. (1991). *Essentials of Behavioral Research: Methods and Data Analysis. Second Edition*. McGraw-Hill Publishing Company, pp. 46-65.
- Sawilowsky, S. S. (2000). Psychometrics versus data metrics: Comments on Vacha-Haase's "reliability generalization" method and some EPM editorial policies. *Journal of Educational and Psychological Measurement*, 60, 157-173.
- Segall, D.O (1994). The reliability of linearly equated tests. *Psychometrika* 59, 361 - 375
- Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120.
- Vacha-Haase, T., & Thompson, B. (2010). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development*, 44, 159-168.
- Wells, C.S (2003). *Reliability and Validity*; Available from: <="" a="">. [Last accessed on 2015 Dec 09].
- Wilkinson, G. S., & Robertson, G. J. (2006). *Manual for the Wide-Range Achievement Test (WRAT-4)*. Los Angeles, CA: Western Psychological Services.
- Wong, K.L, Ong, S.F,& Kuek, T.Y (2012). Constructing a survey questionnaire to collect data on service quality of business academics. *European Journal of Social Science*. 29:209-21.