Use and abuse of reliability in research: An analysis of postgraduate theses at Catholic University of Eastern Africa, Kenya

Francis Omillo - Okumu

Dr.Sc. Francis OMILLO - OKUMU

Abstract

Reliability is a classical quality measurement criterion that checks stability, equivalence, internal consistency and absence of errors in the measurement instruments. Due to the need for accurate and reliable results for decision making, a lot of emphasis has been put on proper use of evaluative measurement criteria. Reliability as a criterion of evaluating measurement tools is the focus of this paper. The study aims at 1) finding out the extent to which reliability has been wrongly used; and 2) establishing how Likert type scale has been abused among postgraduates? Classical test theory guided the study to answer the research questions. Descriptive research design was adopted to direct in collection and analysis of data. The study population is theses authored by postgraduate candidates for the years 2018, 2017 and 2016 in Catholic university of Eastern Africa, Gaba Campus. Out of 126 theses filed at the university library, 40 were sampled for the study using the 30% rule. The researchers reviewed literature on research from Loreto library and e-libraries using search engines. Data was collected using review of the sampled theses guided by structured tools developed from the research questions. Collected data was analyzed using descriptive statistics. Study findings revealed that most of the theses (45%) adopted Test-retest, followed by Cronbach (42.4%) as reliability test tests. Test-retest users confused right techniques

ILIRIA International Review – Vol 10, No 1 (2020)

and coefficients to determine and interpret as stability of measurement instrument. Cronbach was largely pinned at 0.7 alpha without proper interpretation of levels of internal consistency. The third most favored techniques was Split half (10%). The academicians who used split-half did not fix the inherent weakness of underestimation. Lastly the study found out that almost all of the theses used a pentatonic Likert-type scale and considered it as an interval scale instead of ordinal scale of measure. The study suggests that teachers of research should strengthen knowledge dissemination on quality measurement in tools of research by observing maximincon principle. Secondly young researcher should be ready to learn new techniques of quality measurement for reliable and accurate research outputs.

Keywords: reliability; use and abuse; postgraduates theses;

1.0 Introduction

1.1 Background to the research problem

Postgraduate theses are not only meant to grade students for graduate, but for consumption by relevant institutions. The institutions use the research outputs to make decisions. The decision making process must be fed with true and reliable information provided by research. Reliability is a classical quality measurement criterion that checks stability, equivalence, internal consistency and absence of errors in the measurement instruments. The concept of reliability descends from the positivism paradigm of research. This paradigm demand research to be scientific; that all empirical observations be measured objectively and verifiably. Therefore, all measurements must demonstrate reliability, validity, measurement scales, sensitivity and operational definitions (Zikmund, Babin, Carr & Griffin, 2010; Pittenger, 2003). Measurement in research gives meaning to data by assigning numbers. The numbers assigned to the observed aspects of a phenomenon must follow established set of mapping rules (Schindler, 2011); or 'operational definitions' that assigns numbers to empirical phenomena under four assumptions or measurement scale: nominal, ordinal, interval and or ratio (Cooper & Schindler, 2011).

Beyond quantifying behavior, measurement scales critically determines data analysis and inference making. Each of the measurement scale has unique characteristics and shows data in special manner. Nominal scale, for example, assigns numbers to things in a mutually exclusive manner. Apart from ordering things alphabetically, it neither ranks nor indicates that one thing is better than the other (Pittenger, 2003). Example of nominal classification are sex (female = 1, male = 2) and religious preference (Atheist = 1, Buddhist = 2, Christian = 3, Hindu = 4, Muslim = 5). Ordinal scale quantifies things by ranking; showing the lowest and the highest e.g. Likert type scale: very good = 5, good = 4 neither good nor bad 3, bad = 2, very bad = 1. Though commonly used by most students, it lacks precision and accuracy. Interval scale assigns number to things in an ordered and ranked way with equal distance between numbers. Example of interval is when estimating the age of enterprise (start up 0-3, growth stage 4-7, maturity 8-11). Ratio scale assigns numbers to things in an ordered manner with equal intervals. Zero (0) in ratio scale is an absolute value and represents lack of construct. In interval scale, zero is an arbitrary and convenient point. Ratio can be used to measure weight, speed, time etc.

Once measurement scale is determined a tool or measurement instrument is developed by coming with constructs and items. The instruments can take a form of clinical simulations, survey questionnaires and tests of attitude, skills and knowledge. It is this instrument that has to be subjected to scrutiny to ensure that it testes what it was meant to test in an accurate way. A measurement instrument is reliable if it produces similar results after repeated observation under identical conditions (Mikkelsen, 2007). It is also reliable when it portrays consistency of measurement (Shaughnessy, Zechmeister, & Zechmeister, 2003), meaning that it is dependable. It is also considered reliable if the ratio of variance in true score influences the observed scale score is high. Stability is the ability to secure consistent results after repeat tests, what Pittenger (2003) calls 'consistency of measurement.' These facts are not clear to many young researchers. More often than not they are confused and misinterpreted.

1.2 Statement of the problem

Ideally postgraduate students form the epitome of academic research in Academia. Their accurate research outputs can help institutions change for better. Yet they have fallen sort of accuracy in their research measurement tests. Most postgraduate students doing masters and doctorate studies fall in problem of using wrong techniques to establish reliability and interpreting it. During my encounter with most students during their postgraduate defenses, I have learnt that reliability has been abused, thus occasioning inaccurate results. The results therefore cannot be dependably used by executives to make right decision. The common causes of inaccuracy are candidates' failure to identify and aligning right reliability measure with measurement technique; confusing thresholds of various reliability techniques; wrongly interpreting outputs of reliability test; and confusing Likert-type of scale to be interval. It is behind this backdrop that research is founded. The outcomes of this study are to help researchers properly test and come up with accurate instruments that would yield reliable results for policy and decision making.

1.3 Research questions

1) To what extent has reliability been wrongly used among postgraduate candidates in Catholic university of Eastern Africa?
2) How has Likert type scale been abused by postgraduate candidates in the Catholic university of Eastern Africa?

2.0 Literature Review

This section critically analyses selected works so as to unveil the known and the unknown as well as the gaps that need filling. It starts with analysis of classical theory on which the study is based, followed by known approaches of reliability and the conceptual framework.

2.1 Theoretical framework

This study is anchored on classical test theory that states that summation of true score (T) and a measurement error is equivalent to the observed score(y). Error would be zero if the measurement would be free of biasness and measurement errors (Eisinga, te Grotenhuis, & Plezer, 2012). In other words a reliable measurement instrument is one that has no error. However, it is almost impossible to have no errors in measurement. It is the work of business researchers to reduce the errors as much as possible using maximincon principle that maximizes the changes in individual differences and minimizes the error variance. As regards reliability measurement errors are common among young academicians that need to be reduced by unveiling critical points of their committal and prescribing requisite measures.

2.2 Reliability estimates

There are five approaches which include: stability, equivalence, internal consistency and absence of errors in the measurement instruments. The five have been broadly grouped into inter-rater and intra-rater. Inter-rater reliability is about the level of agreement of two independent raters in their observation. Intra- rater reliability involves one rater or observer's consistency e.g. internal consistency.

2.2.1 Stability of measurement instrument

It is also called stability evaluation of research instruments. Stability refers to obtaining consistent re results from the same subject using the same instrument severally. An instrument is stable if it secures consistent results over time on the same person repeated severally. This approach is concerned with consistence over time and reproducibility of measurement (Oliveira, Santos, Carvalho, & de Araujo, 2016). Researchers and statisticians have identified *test-retest* a good estimator of stability because it compares and correlates the findings of two or more tests on the one person in more than one occasion. The correlation between two or more test scores using the same instrument at different time intervals yields a *coefficient of stability* which is the information about the stability of characteristics across different occasions (Cronbach, 1951). This estimator is highly determined by the type and purpose of measurement. In circumstances, where attrition is high or the subject is undergoing drastic changes, stability evaluation is not recommended. On one hand Zikmund, Babin, Carr and Griffin (2010) find the evaluation is about *repeatability*, on the other hand Oliveira, Santos, Carvalho and de Araujo, (2016) find it about *reproducibility*. Using test-retest, Shaughnessy, Zechmeister and Zechmeister (2003) afers that the threshold is 0.8; meaning that coefficient of stability, repeatability or reproducibility is desirable when it is at 0.8 and above.

2.2.2 Equivalence of measurement instrument

Two forms that are either equivalent or parallel but not identical are used to test variability of results. The procedures include developing *two parallel or equivalent forms* and make each person should respond to the items in the two forms (meaning each respondent will have a pair of results). Finally, compute a *coefficient of equivalence* using a correlation formula (Magnusson, 1967). Coefficient of equivalence gives information on the extent to which different sets of items intended to measure same characteristics can be interchanged. This estimator is recommended where there is high attrition and memory lapse. However, the trouble of making two forms, rigor of computing means and standard deviations in each of the two score make it tedious and quite demanding. Another demerit is the fatigue and boredom respondents end up with due to being subjected to two tests. It is requires longer time, too.

2.2.3 Internal consistence of measurement instrument

A measurement instrument is internally consistent when all items in it measure the same construct. It is the degree to which items within a measurement tool are inter-relatedness. *Coefficient of consistency* does not show homogeneity of items in a measurement tool. According to Tavakol and Dennick (2011), internal consistency is a necessary but not sufficient condition for dertemining homogeniety among sampled items. The tests of internal consistency include: *split-half reliability, Cronbach coefficient alpha* and *Kuder-Richardson formulas 20 and 21*(KR20, KR21). Split-half indicates the level of error in a test score due to mistakes in construction. In split-half, the researcher picks results of half of the items and compares them with the other half using odd versus even numbered items. Then compute totals for each half. Last is to correlate the scores using standard formula. A tool will internally consistent if two halves of the scale yield similar results and highly correlate. Weakness of underestimating the true reliability because of dealing with halves has been observed in split-half. Therefore, Spearman-Brown prophecy formula, Rulon formula or Guttman formula are recommended to enhance split-half reliability because they adjust length effects and measures coefficient of consistency of the whole test (Kerlinger & Lee, 2000).

Spearman-Brown Prophecy formula was developed by a British psychometrician Charles Edward Spearman (1863-1945) to evaluate measurement tools' internal consistency based on split-half findings by converting half-length estimates to full length estimates. Also called Spearman-Brown Prediction formula or Correction, is used to correct split-half findings which are limited to half the test to address the reliability of the full length test in a two-item scale (Eisinga, te Grotenhuis, & Plezer, 2012). It takes split-half coefficient as input to yield a full-length test coefficient.

$$\rho^*_{xx'} = \frac{n\rho_{xx'}}{1+(n-1)\rho_{xx'}}$$

where pxx′ is the split half reliability; n is the number of items; and p*xx′ is the full-length equivalent.

Rulon implies that consistency in total test score is the proportion of true variance in a measurement tool. It has also been proved that Rulon formula can apply in parallel forms tests, where coefficient of equivalence would be composed of the sum of the two forms (Anastasi, 1982).When applying Rulon formula for computing Split-half, the researcher need to divide the variance of the differences between each person's scores on the two half tests by the variance of the total scores and then minus the answer from 1.

$$r_{tt} = 1 - \frac{\sigma^2 d}{\sigma^2 t}$$

Where d is the difference between two half scores of a respondent; d = SD of those variances; and t = SD of total scores.

Last is Guttman formula. Guttman proposed 6 measures of evaluating consistency in tests, all using the lower bounds. To fix the weakness of underestimation in split-half test, Guttman prefers Lambda 4 which inserts the covariance between the total of two groups of items on the mean of the variances to generate Guttman split-half coefficient. The formula is stated as follow:

$$\lambda = \frac{4cov(h_1, h_2)}{var(t)}$$

Where h_1 is the partial scores from the first half; h_2 is the partial scores for the 2nd half; t is the sum of scores and λ is the Guttman's maximum split-half coefficient produced of all possible halves. Though Guttman Lambda 4 is recommended as a better measure for reliability, it suffers from two weaknesses. First, for this formula to be used, the variances of the two splits should be equal. Secondly, Guttman is likely to over-estimate when the sample size is small and also there are many items in the instrument. When using split-half tests, the guide for interpreting the coefficient of consistency is as follow: between 0.8 and 0.95 it is very good;

between 0.7 and 0.8 it is good, 0.6 and 0.7 is fair and below 0.6 is poor (Zikmund, Babin, Carr, & Griffin, 2010).

In 1937, Kuder and Richardson developed formulas (KR-20, KR-21) based on understanding that mean and variance are the same for every item in the test. The two versions; that is, the twentieth equation 20 and KR-21 the improved version tests item difficulties or endorsement. The test is recommended for binary scoring tests. Kuder-Richardson formulas 20 and 21(KR20, KR21) can't apply where the tool is not formatted in a dichotomous style.

Another test that is equivalent to split-half test for internal consistency of measurement tool is *Cronbach coefficient alpha.* The test was developed by Lee Cronbach in 1951 as a means of all split-half coefficients using a formula developed by Guttmann in 1945. It is also called *tau-equivalent reliability* test is a lower bound estimate, too. The formula is as follow:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

Where: N is the number of items; $\bar{c}$ is the mean covariance between item pairs; and $\bar{v}$ is the mean variance.

Initially, the coefficient alpha was developed as a test for internal consistency for psychometric instruments. Zikmund, Babin, Carr and Griffin (2010) observed that the coefficient alpha is more frequently used by researchers in multiple- scale instruments, most of which use likert type scale. The Cronbach coefficient alpha indicates the level of inter-item correlation and their relatedness. It is a convergence measure, that is; it tests whether the different items converge on one construct. The computed alpha shall always fall between 0 and 1 value. Values tending towards one should be considered as internal consistency of measurement instrument being high. Values toward zero indicate poor internal consistency of measurement instrument. This rule of the thumb should be applied with caution when interpreting the computed coefficient alpha. Specifically the values can be interpreted in bracket as follow (DeVellis, 2012):

Table 2.1: Cronbach Values Interpretation

Value	Interpretation
Less than 0.6	Poor consistency
Between 0.60 and 0.65	Undesirable
Between 0.65 and 0.70	Minimally acceptable
Between 0.70 and 0.80	Respectable
Between 0.80 and 0.90	Very good consistency
Above 0.90	Unacceptable

Compiled from DeVellis (2012)

Most researchers falsely consider values above 0.90 to be more correct to mean the instrument is more internally consistent. In fact they consider anything above 0.7 to be good. According to DeVellis (2012), when the cronbach coefficient alpha is above 0.90; it means that items are highly correlated. Therefore the test suffers from multicollinearity and the researcher should reduce items in the scale.

Other uses of coefficient alpha are dimensionality and sample size determination. According to Bujang, Omar and Baharum (2018), the coefficient alpha measures dimentionality or homogeneity. When calculated value tend towards one(1), it indicate that items are measuring the same dimention(homogeneous) and when it tends towards zero(0) value, it indicates that the items are measuring towards different dimentions(heterogeneous). This interpretation is hotly contested by other scholars. Homogeneity concerns single latent construct and is measured by factor analysis (Tavakol & Dennick, 2011). Hoekstra, Vugteveen, Warrens and Kruyen (2018), equally observed that it would be an abuse of coefficient alpha for it to determine dimensionality. Of late Cronbach alpha has been stretched to determine sample size using excels software and sample size tables. For example, in a study reviewing sample size determination, it was found that a larger sample size was achieved when Cronbach coefficient alpha was set at 0.5 instead of 0.7 (Bujang, Omar, & Baharum, 2018). However, the Coefficient alpha has been criticized for not being of value when evaluating short measurement tools and those that cannot be split into distinct subsets (Hoekstra, Vugteveen, Warrens, & Kruyen, 2018).

2.2.4 Accuracy of measurement tools

Another perspective of reliability is the check on absence of errors in the measurement instruments. An error is an event that either adds or subtracts a score every time a test is in use. Measurement being the core of positivist and scientific studies always suffer from measurement errors. Causes of errors that affect reliability are as shown in the table 2.2 below.

Table 2.2: Sources of measurement errors

Type of error	Conditions of occurrence
Instrument error	When instrument stops working; poor wording of questions; test not being precise.
Respondents' variability	When participants suffers from fatigue and boredom; participants not heeding instructions; uncooperative respondents; and respondents not answering some questions.
Researcher's variability	Poor recording of data; failing to consistently follow procedures
Environmental variability	Disturbances from environment e.g. noise; respondent's discomfort; difference in conditions of measurement for various respondents

Source: Pittenger (2003)

For the test to be reliable, the researcher should minimize the errors as much as possible. It begins with finding out the causes of errors in the measurement plan and crafting steps to alleviate them.

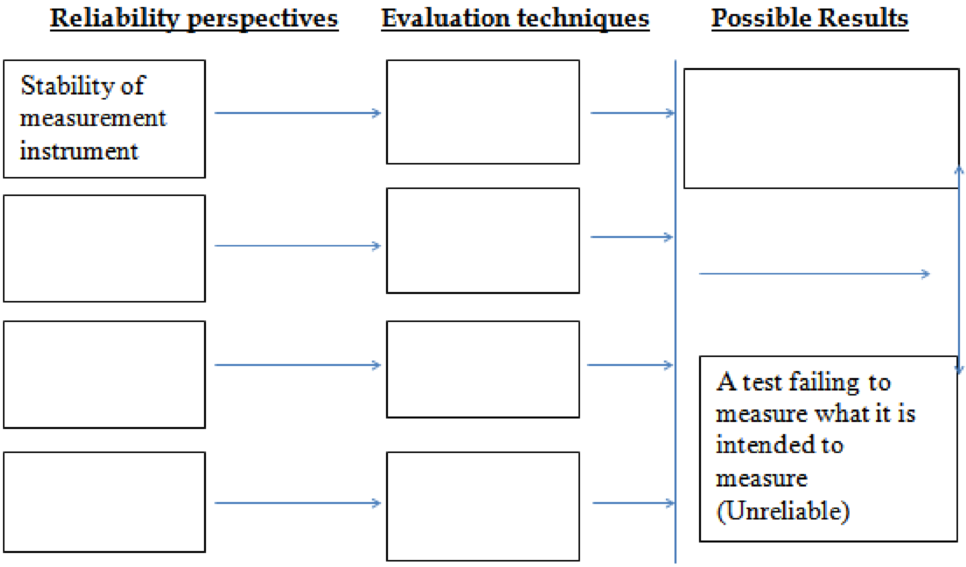
How is accuracy measured in measurement instruments? Ideally, accuracy in test can be obtained through correlating scores on the same instrument administered twice on the same group of participants. However, this approach is hypothetical and not practical (Tavakol & Dennick, 2011). A better alternative is using interrater reliability techniques. Inter-rater reliability is about the level of agreement of two independent raters in their observation. It is obtained by correlating different raters or observers e.g. equivalence. When the inter-rater reliability is high then the researcher can be sure of accuracy in the test measuring what it intends to measure. It is determined by calculating a percentage agreement or correlation.

$$\text{Inter-rater/observer reliability} = \frac{\text{Number of items agreed by observers}}{\text{Number of opportunities to agree}} \times 100$$

The outputs of this formula are *coefficient of precision*. It is the degree to which the instrument is accurately measuring what it is intended to measure.

2.3 Conceptual Framework

Figure 2.1: Conceptual framework of reliability, evaluation techniques and results


Reliability perspectives
Evaluation techniques
Possible Results
Stability of measurement instrument
A test failing to measure what it is intended to measure (Unreliable)

3. Research Design and methodology

The study adopted descriptive design. The study population was postgraduate theses at the Catholic university of Eastern Africa filed at the Loreto library, Gaba Campus. The Campus is in Eldoret along Eldoret-Kisumu Road. The study population covered three cohorts: 2018, 2017 and 2016. According to the sampling frame from the Loreto library, there 126 theses filed, which formed the study population. The theses were a partial fulfillment for the requirements in postgraduate diploma in education, masters in education, masters of Arts in development studies, masters in business administration and doctor of philosophy in education.

Stratified random sampling was calculated using the 30% formula as shown below.

Table 3.1: Stratified sampling of research students

COHORT	Population	30% rule	Sample
YEAR	N	n	Adjusted n
2018	31	9.3	10
2017	58	17.4	18
2016	37	11.1	12
TOTAL	126	37.8	40

Source: Loreto Library data
Compiled by Author (2020)

Table 3.1 shows that out of population of 126, a sample of 40 theses was arrived at for study. The highest number of sample (18) was from the 2017 cohort because it hard the largest population of 58 theses. After sampling, data was collected by desk reviews using structured study guide based on research questions. The source of data was secondary from the same library and online libraries using computer search engines. The data was from sampled theses, books and peer reviewed journals on reliability. The collected data was analyzed using descriptively. Descriptive statistics captured data and presented them through percentages, means, standard deviation and frequencies. The information was displayed in bar charts, graphs and pie charts. Content analysis of data from the theses was also used to explain the interpretations of reliability and evaluation techniques employed.

4.0 Research Findings and Discussion

4.1 Background information

This section considers demographics of the authors of the theses filed at Loreto Library, at Gaba Campus. Demographics, particularly gender is considered a critical variable to the study because it informs on levels of scholarship differences at individual and the university. The findings of the study show that women-authored theses were the majority 24(60%) as compared to male-authored theses 16(40%). The implications are that there more women completing postgraduate studies than men. It can also be adduced that women are more daring to research than men. The findings confirm similar results of a comparative survey on gender differences in research scholarship among academics where proportion of women in postgraduate studies had risen to significant levels (Jung, 2016). In United States of America, the proportion women with advanced degree above bachelors had outnumbers men by 1980s, though not in fields of engineering, mathematics, computer sciences and physics (Hyde, 2014). It means that the old trend of male-skewed research is diminishing and women are catching in academic circles.

After determining gender, the study sought to determine the degrees and their popularity at Gaba Campus. The study found that there were five postgraduate programmes: postgraduate diploma in education (PGDE), Master of Arts in development studies (MA), Master in Business Administration (MBA), Masters in Education (MED) and Doctor of Philosophy in Education (PhD). The most popular is Master in Business administration 28(70%), followed by masters of Arts in development projects6 (15%) and MED 3(7.5%). This implies that there is greater demand for masters of business and masters in development studies in the Kenyan market.

4.2 Usage and abuse of reliability tests

On the usage of evaluation techniques majority used test-retest18 (45%), followed by Cronbach 17(42.5%), Split-half 4 (10%) and others 2.5% as shown in table 4.1. Using test retest implied that the researchers were after stability, repeatability and reproducibility of the instrument as an aspect of reliability. Not internal consistent.

Table 4.1: Reliability techniques used by postgraduates

Reliability type	Technique of measure	Frequency	Percentage
Split-half	Cronbach alpha; Pearson product; spearman Brown prophesy; Nil	4	10
Test-retest	Cronbach alpha; Pearson product moment; interrater; Pearson correlation; Nil	18	45
Cronbach	Cronbach; nil	17	42.5
others	nil; document analysis, expert	1	2.5
Total		40	100

Source: Loreto Library
Compiled by Author (2020)

The researchers who used this method subjected same respondents twice to the interviews at close intervals of one to two weeks, according the theses reviewed. The study revealed that only 3 theses (7.5%) got right score, above 0.8 recommended coefficient of stability (Shaughnessy, Zechmeister & Zechmeister, 2003). This implied that 92.5% of the instruments that were used in the theses lacked requisite stability, repeatability and reproducibility. Out of the 40 theses analyzed theses, 12.5% did not compute and show the coefficient of stability. This means that the candidates found difficult to compare and correlate the findings of two or more tests on the one person in more than one occasion.

On Cronbach, all theses reviewed considered 0.7 to be the threshold of reliability and anything above it was reliable which is false. According to DeVellis (2012), Cronbach alpha determines internal consistency dimension of reliability. The alpha range from 0 to 1, from poor consistency < 0.6 to unacceptable when > 0.9. Most researchers unwittingly clamp together coefficient of above 0.7 to be good. Zikmund, Babin, Carr and Griffin (2010) recommends cronbach should be interpreted as the level of inter-item correlation and their relatedness. The technique is best in multiple- scale instruments.

Split-half was found not to be as popular as the first two among the postgraduates. About 10% prefered it to the other techniques of testing reliability. Despite split half testing internal consistense all the candidates who used it either interpreted it as a measure of reularity or consistency of results. The second mistake is that all the users did not consider the weaknes inherent in the method and consequently address it. Split-half underestimates true reliability. To fix the weakness, Spearman-Brown prophecy formula, Rulon formula or Guttman formula are recommended to adjust the length effects of the whole test (Kerlinger & Lee, 2000).

Table 4.2 summarizes corrections as gotten from theses analysis as follow:

Table 4.2: Corrections of Interpretations of Reliability

	Usage	**Interpretation by candidates**	**Correct**	
			Interpretation	**Techniques**
1.	Split half (10%)	Regularity of measurement over time and conditions; consistency of results; trust worth	*Internal consistency of measurement instruments: Cronbach alpha rules applied as below.*	*Correlation; Spearman-Brown prophecy formula; Rulon formula; or Guttman formula*

2.	Cronbach (43%)	Internal consistency; degree of consistent results after repeat trial; consistency of instrument results; degree of consistency in results; consistency of measurement; possibility of instruments yielding similar results; consistent measure of what is supposed to be measured; reliable results	*Internal consistency of measurement instruments:*
			< 0.6 is Poor consistency;
			Between 0.60 & 0.65 mean Undesirable;
			Between 0.65 & 0.70 is minimally acceptable;
			Between 0.70 & 0.80 is respectable;
			Between 0.80 & 0.90 means Very good consistency
			Above 0.90 is Unacceptable
				Cronbach coefficient alpha:

3.	Test-Retest (45%)	Reliability of instrument; precision and accuracy; stability of instrument; reliable results; consistency in measuring procedures that produce similar results over a number of repeat trials; reliability and consistency; degree of consistency	*Stability of measurement instrument: threshold is 0.8; meaning that coefficient of stability, repeatability or reproducibility is desirable when it is at 0.8 and above*	*Correlation: coefficient of stability*

Source: Author (2020).

According to table 4.2 the three most popular tests for reliability are given proper guidelines for interpretation and right techniques to compute reliability. For example tests retest threshold bar for stability is 0.8 which gotten through correlation techniques. Cronbach is no longer determined by the 0.7 threshold, but described according to the six classifications of alpha, ranging from poor consistency to unacceptable levels. Split-half, according to the table, is never complete with correlation alone. It is supposed to be enhanced with Spearman-Brown prophecy formula; Rulon formula; or Guttman formula to address the inherent weaknesses.

4.3 usage and abuse of Likert type scale

The study found that 95% of the postgraduate candidates used Likert-type scale and only 5% applied document analysis. They graded psychological attitudes of subjects using either on four-point (2.4%) or original pentatonic (92.5%) scales as shown in table 4.3.

Table 4.3: Usage of metric scale

Type of instrument	No.	%
pentatonic Likert	37	92.5
4 point Likert scale	1	2.5
Document analysis	2	5
Total	40	100

Source: Loreto Library
Compiled by Author (2020)

The 4 point rating system had strongly agree (SA), agree (A), disagree (DA) and strongly disagree (SD) options. The pentatonic included the neutral (N) option. Despite the fact that this psychometric gradation is most loved by the postgraduates, the system is highly discouraged of late due to its inability to capture the intended constructs (Luciano, 2016). Specifically, the scale suffered from 'social desirability bias,' where respondents tended to answer questions in conformity with social expectations (Subedi, 2016). Also there was 'response bias' observed in Likert-type scale. This is a tendency of respondents failing to be objective and trying to score the middle throughout by avoiding the extremes. In other words the traditional Likert type of gradation is obsolete and inappropriate. Therefore researchers have to rethink the use of it and adopt improved psychometric scales.

The candidates who used the Likert type of scale confused it to be interval scale and furthered to use it in parametric techniques like linear regression for analysis of data. Likert type scale is ordinal. It does not have equal distances between options and therefore not compatible with parametric techniques (Dewinter & Dodu, 2010). Confusion and misinterpretation, such as these, lead to invalid and inaccurate research outputs that cannot be relied on by executives for proper decision making.

5. Conclusion and suggestions

5.1 Conclusion

Demands of information age require that research outputs should adhere to classical test theory principles so as to be free from error for proper decision making. This can be achieved through reliable measurement tests. According to the study most young researchers are still far from it. The young researchers often use test-retest, Cronbach and split-half which they have failed to observe rules governing them and their use. The study also found out that almost all of the theses used the old pentatonic Likert-type scale and considered it as an interval scale instead of ordinal scale of measure. This breeds errors in scientific research that can lead to poor decision-making process.

5.2 Suggestions for improving reliability of the measurement tools

How can students improve reliability in measurement tools? According to Kerlinger and Lee (2000) reliability of tests can be enhanced through *maximincon principle* that is, reducing error variance and increasing variance in specific differences. Specifically the principle demands that;

a) Students of business research write items in the test clearly to avoid possibilities of interpreting items in more than one way.
b) Inventing items and refining questions (Blumberg, Cooper, & Schindler, 2011); that is increasing relevant items of the same kind and quality in the test.
c) Standardize the test by applying simple to understand standard instructions for admiration and scoring of measurement tools.

List of References

Anastasi, A. (1982). *Psychological testing.* New York: Macmillan.

Blumberg, B., Cooper, D. R., & Schindler, P. S. (2011). *Business research methods.* London: McGraw-Hill Higher Education.

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach's alpha test: A simple guide for researchers. *Malaysian Journal of Medical Science*, 85-99.

Cooper, D. R., & Schindler, P. S. (2011). *Business research methods.* New York: McGraw Hill International.

Cronbach, L. J. (1951). Coefficient alpha and the internal istucture of tests. *Psychometrika*, 297-334.

DeVellis, R. F. (2012). *Scale development .* California: Sage Publications.

Dewinter, J., & Dodu, D. (2010). Five point Likert item: t test versus mann-whitney-wilcoxon. . *Practical Assessment, Research & Evaluation*.

Eisinga, R., te Grotenhuis, M., & Plezer, B. (2012). The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown. *International Journal of Public Health*, 1-13.

Hoekstra, R., Vugteveen, J., Warrens, M. J., & Kruyen, P. M. (2018). An empirical analysis of alleged misunderstandings of coefficient alpha. *International Journal of Social Research Methodology*, 351-364.

Hyde, S. (2014). Gender similarities and differences. *Annual Review of Psychology*, 373-398.

Jung, J. (2016). Gender differences in research scholarship among academics: An International comperative perspective. *ResearchGate*, 163-178.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundation of behavioural research.* Belmont: Cengage Learning.

Luciano, R. (2016). Rethinking the use of Likert scale: Tradition or technical choice? *Brazilian Journal of Marketing 9(1)*, 11-26.

Magnusson, D. (1967). *Test Theory.* Reading, MA: Addison-Wesley.

Mikkelsen, B. (2007). *Methods for development work and research.* New Delhi: Sage Publications.

Oliveira, K., Santos, B., Carvalho, F. M., & de Araujo, T. M. (2016). Internal consistency of the sel-reporting questionnaire. *Revista de Saude Publica*, 1-10.

Pittenger, D. J. (2003). *Behavioural research design and analysis.* Boston: McGraw Hill.

Schindler, D. R. (2011). *Business research methods.* New York: McGraw-Hill International.

Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2003). *Research methods in psychology.* Boston: McGraw Hill.

Subedi, P. (2016). Using Likert type data in social research: Confusion, issues and challenges. *International Journal of Contemporary Applied Sciences*, 36-49.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 53-55.

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2010). *Business research methods.* London: South-Western Cengage Learning.