CONJUGAL RELATIONSHIPS BETWEEN ASSESSMENT AND RESEARCH

H. Johnson Nenty, PhD

Abstract

While educational assessment tries to model that which is latent (i.e. cannot be seen, felt, heard, or even sensed) , in an individual, using the results of a confrontational interaction between the trait or ability involved and appropriate tasks; educational research generally tries to model a population's behaviour, which is also latent, based on what is observed to hold for a representative sample. Hence both assessment and research share a lot in common as inferential processes, which try to find out the truth about individual behaviour, in the case of assessment; and about mostly population behaviour, in the case of research. Three types of variance that influence these processes: systematic desirable variance from the trait or ability under assessment or the independent variable in research which must be maximized; systematic error variance emanating from other extraneous sources in both assessment and research, which must be controlled; and the ever-present random error variance which must be minimized in other to attained reliable results.

Key words: Assessment, research, measurement, error, systematic desirable variance, systematics error variance, random error variance.

Introduction and Definitions

The results from assessment and research are the two most important drivers of policy formulation and implementation especially in developed countries (Pohjola & Tuomisto, 2011; Dhaliwal & Tulloch, n.d.) as well as that which sustain educational decision from classroom to government levels. Both assessment and research provide a means for democratizing policy-making processes as they involve the participation by, and inputs from, several stakeholders and the public. A policy is valid to the extent that inputs into its formulation are valid and since assessment and research are the two most important sources of input into the formulation of education policy, to ensure the formulation of valid policy, we should ensure that the score we arrive at through assessment, and the findings we reached through research are valid. Have we ever sat down to ponder why the implementation of most of our education policies do not always yield solution to the problem for which the policy was developed? More often than not, it is as a result of the

quality of information provided by research and assessment as input into the policy formulation. Assessment and research are partners in crime in the provision of invalid data as input into policy formulation.

In the face of an array of confusion when it comes to the definition of some basic terms involved in this paper, there would be no need to add to such situation by bothering you with more definitions. But since your understanding of my views in the paper depends much on your acquaintance with my unique views on some of these terms, I plead to be allowed to define terms related terms to the presentation. Assessment and research are conjugally interdependent, for example, scientific research cannot do without assessment but assessment have other several uses other than it uses in research. Both assessment and research are processes of searching for truth. In education, while assessment is a systematic process of searching for the truth about the type and amount of a given characteristic or behviour possessed by an individual, research is a systematic process of searching for the truth about a population in the process of finding a solution to a problem or satisfying one's curiosity or interest. Assessment gathers and analyses data with which to infer an individual's standing on a given trait while research gathers and analyzing data with which to infer a population's standing on a given trait. Both try to find out what is not known or well understood – the truth.

Education

To Kerlinger (1986), science is a process as well as a product. If engineers want to understand in order to modify characteristics or form of physical materials they grab the scientific process. Similarly if medical doctors want to understand and ensure desirable changes in the characteristics of human health they grab the scientific process. Education, which is seen as the process involved in ensuring and maximizing desirable changes in human behavior, like other professions, has no science of its own but according to Brubacher (1939), "like medicine, education science is based on other sciences" (p. 15). Education science adapts the scientific process in her attempt to study and understand human behavior and hence create knowledge about human behavior. Equipped with a good knowledge and understanding of human behavior, educators can do a better job at trying to change it desirably. The two most important aspects of the scientific process are measurement and research. This put assessment and research at the center of every effort of the education process. Any serious intention to improve education must necessarily give assessment and research a prime of place.

As a process of manipulating human and environmental resources directed at provoking desirable changes in learners' behavior education consists of three components: the inputs; the processes and the products. The inputs are all the human and material resources made available to education; the processes are the actions taken to arrange and manipulate the interaction among the resources and the behavior of the learners; and the

outputs are the results from these interactions as regards mainly the desired changes in the learners' cognitive, affective and psychomotor behavior.

Assessment

Assessment is concern with the quantity and quality of all the input into and processes involved in education and their outcomesIt is through assessment that education is defined, its processes monitored and its products documented. The quantity and quality of the 'change' is determined by assessment, and how much 'change' is acceptable for one to be said to have 'learned' is determined by evaluation. Feedback from assessment enhances teaching and learning. Assessment determines who is qualified to be admitted for education at each level of the process and determines and ensures that some acceptable standards are met at each stage for acceptable progress. The inputs into the education process are checked for quality, the processes involved in education are monitored and improved and the product of education is determined, quantified, and certified. Assessment determines the desirability of the changes, the amount of such changes and validates the meaning as well as the processes involve in and that bring about the desirable changes. Assessment inputs into decision-making processes from the classroom level, through the homes, to government levels. In other words assessment collects information emanating from the input, processes and outputs of education, analyses them, and feeds the results into decision-making processes from the classroom to government levels and into research, especially evaluation processes. To Nenty (1997a), assessment as applied to education, is:

> anything done to find out what knowledge, skills, habits, attitudes, practices or generally what behaviour a learner does or does not have, acquire, or develop, before, during, and at the end of an instruction, a period of instructions, or a course of study . . . . The "anything done", includes: observing, interviewing, professional experience/judging; using questionnaire; classroom questioning; project assignment; class or seat-work; homework assignment; classroom testing; measuring; examination, etc. (par. 4).

Hence assessment consists of several processes in education and inputs into research, evaluation and policy- and decision-making from the classroom to government levels (Fig.1).

The assessment process involves:

1. Theory-based conceptualization of the trait, ability or behaviour to be assessed.
2. Generation of several indicators/indicants of the conceptualized trait or behaviour.
3. Converting indicators/indicants into provocative cognitive, affective or psychomotor tasks guided by the objective of the assessment.
4. Confronting an individual's trait or behaviour under assessment with a good number of these tasks.

5. Converting the result of the trait-by-task interaction into observable quantity or quality.
6. Using resulting numbers to estimate the quantity or quality of trait or behaviour under assessment possessed by each individual.
7. Feeding these into evaluation, decision-making or research (see Figure 1).

Figure 1. Assessment as input into evaluation, policy- and decision-making and research

In education, where the product of assessment serves as input into policy decisions, assessment generates and channels the input from learners in the form of scores or grades into the process of formulating and implementing educational policy and practices. In other words, the score produced by assessment takes a central stage where policy is being developed and implemented. When the scores are tumbling, policy are enacted to prop them up and where policy is wanting, valid scores provide a helping hand, as an input into the development of valid policy. Government and the public, has a lot of faith on scores as being a valid indicators of the quality of learning, hence of teaching and education. Many governments in Africa, for example, Botswana, are ready to do what it takes to improve the score. Several studies are undertaken to improve the score. Policies are developed to ensure improvement of the score. Millions of dollars are spent to improve the score. Not only to ensure that every child has the opportunity to earn a score but also to maximize his/her score. There is none of education processes that takes more of human

thinking and produces more controversy than assessment. It has also produced several ways of looking at the best means of reaching its aims. Some of these are discussed below.

Assessment is a means through which concepts and variables are operationalized during research. Operationalization is the process through which constructs, concepts and variables in a research setting are replaced by numbers which are analysed statistically to answer research questions and test hypotheses. Evaluation as a type of research design is appropriate for a systematic objective-driven search for the truth about, and the merit of a programme, project or for example, a curriculum. It is applied in nature and designed to find out which objectives a well-defined programme or project has been met and how well they have been so met (see Fig. 1).

Besides serving as means of generating feedback information, at the classroom level, the result from assessment feeds into evaluation decision based on learners' performance; strength, weaknesses and for remediation determination, behavior and achievement at the cognitive, affective and psychomotor levels. Evaluation interprets or reads meaning into the results of assessment in the light of the prevailing value and standard. It is a value judgment made on an objectively generated information. That is, while measurement as a tool for assessment can be said to be an objective process, evaluation, being a value judgment, cannot be objective as its result depends on which value or standard is used as a benchmark to such judgment. Hence for example, an objectively generated score of 60% for a learner in mathematics, can earn him/her a B, C, D or even a failing grade, depending on the standard based on which his/her performance is judged. Again while john with a height of 1.5m in height might be judged to be a short person among the Fulanis, he might be judged to be tall person among the pigmies but his height, the result of measurement, remains the same.

Assessment for learning is that which aims beyond performance into what learning itself is, and how it can be improved. How can we generate information with which to enhance learning by assessing what learning is? How best learning could take place? Assessing the fundamental process of learning, the effectiveness of each learning activity, that is, the intrinsic meaning of and hence the operationalization of each components of learning. Provides information not merely for ensuring improved performance but also for ensuring improved learning. Eventually it brings about improved in performance much more than assessment of learning does. Assessment of learning, on the other hand, assesses learning given its extrinsic nature, that is, that which shows that learning has taken place. Unlike assessment for learning

which is learning-based, assessment of learning is performance-based assessment. Any type of assessment that 'forms' and empowers a learner for success by ensuring improved performance is formative assessment and both continuous assessment and assessment for learning are aspects of formative assessment. Both summative, and to some extent, continuous assessments are assessment gear to documenting the amount of learning that

has taken place, but assessment for learning is assessment to provoke, ensure, and maximize learning. Summative assessment determines the amount of learning that has taken place and based on its result a terminal decision about the learner is taken. Hence it deals exclusively with performance and not with learning. Whether a person learns or does not learn as long as he/she performs the aim of summative assessment is met. Formative assessment, on the other hand, deals with results or products at the process stage of learning by determining and analyzing the amount of learning that has taken place at that stage and feeding the results back into the teaching/learning process to enhance or improve these activities. Both are external to learning, or are extrinsic to the process of learning, assessment for learning is intrinsic to the learning process and its effect is enhancement of more learning, whetting appetite for more learning, even learning beyond that which summative and formative assessment are concerned with (Nenty & Lusweti, 2015).

Measurement

Among the several means of assessment as listed previously, educational measurement is the most technical of all its tools because it tries to quantify objectively that which cannot be seen, heard, felt, touched or perceived. Hence it is indirect, or like research, it is inferential in nature. Being inferential, it is theory-based, and calls for scientific conceptualization and operationalization of constructs, trait or ability to be measured. For achievement testing, it demands the construction of a domain for the subject matter content (conceptualization of the curriculum contents), and another for the cognitive behavior whose development was intended by the curriculum. Educational measurement is generally accepted as a process, an objective process of assigning numerals to the type or amount of a characteristic or behaviour, which are latent, possessed by a person, a thing or an event.

Unlike physical characteristics, behavioural characteristics are latent and cannot be observed or measured directly as physical characteristics could. Ability, for example, could not be directly measured but could only be inferred from that which results from measuring it indirectly. Ability, for example, is latently inherent in the body that possesses it and to measure it, it must first be provoked or challenged to show up. Since it is latent, through this method, what is actually measured is how much of it is exhibited which might not be the same with how much of it is possessed. Hence to maximize its exhibition, task or challenges that are highly provocative of that particular trait or behavior must be skillfully designed, calibrated and used. Hence, the measurement process is efficient to the extent that it presents challenges, in the form of tasks that are highly provocative of the trait or ability under measurement. In measurement that involves cognitive behavior, such task come in the form of questions or statements and are often called items. An item therefore is a task constructed, validated and calibrated as

a stimulus with which when a testee is confronted would provoke from him/her the amount of the ability under measurement he/she possesses. Each of such tasks has some level of ability-demand, that is, the level of ability under measurement just necessary to overcome the task. Some has it more or less than the others. If such ability is called delta *(δ)* and the ability of the testee is called theta *(θ)* then whether a person overcomes a task depends on how larger his/her theta *(θ)* is than the task's delta (*δ)*. That is, the probability of a correct response to any item in a test depends on the value of *(θ -δ)*. There are some important assumptions that must be met before this holds. So test-taking is a confrontational exercise or an interaction between theta and delta. Following from these, a test as an educational measurement instrument, is a collection of a set of tasks (can be cognitive, affective or psychomotor), often called items, constructed, validated and calibrated as stimuli for the confrontational provocation of the level of ability or trait under measurement which the testee possesses. The results from measurement enables a deep understanding of both the trait being measurement and the items developed to measure it. According to Lord Kevin (1883)

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be (par. 9).

That which results from the measurement are type or amount of what is being measured. They have two important uses. They have no values (not numerical) assigned to them by the process of measurement; values are only assigned to them through evaluation, another important process in education. Evaluation is the process of assigning descriptive (non-numerical) values to the products of assessment. Assessment ends up with a score, for example, of 68% in a mathematics test, while evaluation decides whether that is a good or poor score, represents a pass or failing performance, is categorized as a B or C grade, implies a good or poor progress, etc.

The second consumer of the result of assessment besides evaluation is research. Research deals mostly with variables and variables are characteristics that vary across persons, events, time, etc. Research analyses the unevaluated results of measurement to answer research questions and test null hypotheses. It determines the amount and type of variability among research subjects, as well as the differences between or among groups on a given behaviour, or the relationship between measures, or the level of dependence of one variable on the other. All based on the results of measurement. Hence one can say that without measurement there would not be scientific research. In fact according to Lord Kevin (1883), "to measure is to know. . . . the first essential step in the direction of learning any subject is to find principles of numerical reckoning and practicable methods for measuring some quality connected with it. . . ." (par. 9).In other words, when involved in any scientific process,without measurement,we cannot know or attain the truth.

Measurement Error

Random error of measurement arises when we find that for a true score ($\mathbf{X}_\infty$) that is invariant we observe variability in our observed score ($\mathbf{X}_o$). That is, the same measure taken more than once, yields different scores. It indeterminately occurs whenever physical or psychological measurement is made. According to classical test theory, the score made by a testee ($\mathbf{X}_o$) is made up of a score ($\mathbf{X}_\infty$) emanating from, and thus representing that which is being measured and a score emanating from random error ($\mathbf{X}_e$). That is:

$$\mathbf{X_o} = \mathbf{X_\infty} + \mathbf{X_e} \qquad (1)$$

This classical test theory (CTT) representation does not take cognizance of the presence of error, other than random, which emanates from the actions of factors other than the trait under measurement but which affect the observed score besides the trait we were trying to measure. In actual fact, there are therefore two types of measurement error: random and systematic error. While random error is that which makes a testee's score (Xo) unrepeatable, for example, any variation in performance across examinees and across time; systematic error is that which makes Xo predictably different from what it would have been without it. For example, language in a mathematics test, group belonginess, etc. Given a the same amount of ability, which an examinee or a group of examinees might share with others, anything that makes him/her/them differ in performance is a source of systematic error. Differential item functioning constitute influence of systematic error. Persistently faulty instrumental or human observation or measurement is a good source of systematic error. Considering this thinking, the classical formula in variance form can be re-expressed as (Biesheuvel, 1974; Nenty, 2000; Trochim, 2006):

$$\mathbf{V_o} = \mathbf{V_{com}} + \mathbf{V_{systematic\ extraneous}} + \mathbf{V_e}^2$$

That is, our test score variance is made up of the systematic variance due to the ability our test was designed to measure plus another systematic variance due to extraneous sources; plus the ever-present random error variance.

This provides the basis for the development of item response theory which, based on its unidimensionality assumption, short-chains the variability from all other sources except from that which the test was designed to measure.

Since sources of non-random error brings about systematic not random variation of observed score along with that introduced by the ability under measurement, it is a part of what CTT erroneously calls 'true' score.

While random error makes the distribution of test scores to vary or change across time, occasion, across items; systematic error makes the score to be bigger or smaller than

what it would have actually been without it. Hence random error affects the differences in variability not necessarily the difference in the mean of a given distribution, whereas systematic error affects the mean and not necessarily the variability of test scores (Trochim, 2006). Large random error impedes the reliability of a test whereas large systematic error impedes test validity.

If it were possible to administer a test that measures a specified ability to the same testee repeatedly for say 100 times, erasing his/her memory after each testing, we will generate 100 independent scores representing the same thing. These scores, for no systematic reason, will not be exactly the same, they will differ fluctuatingly around their mean. The mean of these scores will be a good estimate of the true ability of the testee (see Equation 1). The more the number of such measures, the nearer their mean is to the true ability of the testee.If from each observed score this mean is subtracted, the standard deviation of the resulting differences across the 100 measures gives us an estimate of the standard error of measurement (SEM) for the exercise. This is indicative of the random error of measurement

In our everyday practice we measure each testee only once. According to classical test theory, the result of the one-time measure is used as an estimate of the true ability of the testee. Across several testees, we can determine the standard deviation of the differences between each of such scores and their individual scores. The estimated true ability of the testees will differ from the observed score depending on how reliable the test was in measuring what it is measuring. The standard deviation of the scores is used to estimate the standard error of measurement for the exercise thus:

$$S_{measurement} = S\sqrt{1 - r_{xx}}$$

Where $S_{measurement} = S\sqrt{1 - r_{xx}}$ is the standard error of measurement; is the standard error of measurement;

So if we have a test with a reliability of .90 and a standard deviation of 4.660, then the standard error of measurement will be:

$$S_{measurement} = S\sqrt{1 - r_{xx}}$$

$$S_{measurement} = 4.660\sqrt{1 - .90}$$

$$S_{measurement} = 4.660\sqrt{.19}$$

$$S_{measurement} = 4.660\ X\ .436$$

$$S_{measurement} = 2.032$$

This value, 2.032, indicates the amount of random error present in the measurement. The implication of this is that given the probability of .05 error value, any score in the distribution lies within ± 2.032 of its current value. This thinking is always used to justify awarding grades based on interval of scores that result from an examination.

Research

Research is a scientific process of searching for the truth about nature. According to positivist thinking, research does not invent but uncovers the truth which exists independent of human thinking. Hence, there are truths hidden in nature which science is to find out. Since human beings are a part of nature, there are "truths" hidden in each child which education is to "educeree" that is "lead forth" or "bring out" and develop. In other words, there are some truths, in terms of potentials, traits, or generally, behaviour, latent or inherent in every human being which the purpose of education is to explore and then develop (Nenty, 1997b). Finding out the truth about human behaviour is tantamount to creating knowledge of human behaviour, and the process of creating knowledge has been developed and validated through science. According to Brubacher (1939), "like medicine, education science is based on other sciences" (p. 15), it does not have a science of its own. Education science or educational research is therefore, the application of scientific methodology in the search for truth about human nature (Nenty, 1991/92).

While physical sciences study and try to understand and explain the bahviour of the physical or material world, education science studies and tries to understand and explain the world of human behaviour. This it does by using the process of scientific inquiry to study in an attempt to understand, explain, predict and to some extent control human behaviour. This leads to the creation of valid knowledge; and the results serve as input into the development of theories of human behaviour, and provide valid guide and input into the practice and processes of education (Sec. 5, par. 1).

Scientific research is all about the study of variation in characteristic, time, event, observation, experience, behaviour, etc. under the influence of natural or some environmental manipulations. For educational research, it is about variation in human characteristic or behavior, especially in performance. If a characteristic or behaviour does not vary, it is not susceptible to scientific study. The questions are: what causes characteristics or behaviour to vary? Why do they vary for some people more than for others? How can we decrease or increase the variation observed for some characteristics, behavior or performance? A variable is a characteristic or behaviour that varies. If behaviors do not vary, there can be no relationship, nor can there be differences or dependence. Research problem emanates if the relationship between or among two or more variables do not yield desirable results. Effort to solve such problem would call for an attempt to reduce the level of undesirability of such results. To do this, the first thing is to identify the variables involved in the problem situation, determine the degree to which

the variables involve vary, the direction of variability, and the extent to which they vary along or influence each other or why they vary? The undesirability could be eliminated or its level reduced if we can manipulate the influencing variable to change, reduce or alter its influence on the receiving variable. For example, varying teaching method, or level of experience of teachers, etc. to reduce the level of poor performance in mathematics after determining through research that these variables has significant influence on such performance.

Hence research finding is chiefly about how much of the variability of our problem

variable can be accounted for, can be increased or decreased, depending on what was the problem of thestudy, by manipulating the influencing or the independent variable. But while we are trying to joggle the relationship of our influencing variable on our problem variable, the influence of other undesirable extraneous variables have to be taken care of, or we will end up claiming theirinfluences as emanating from our independent variable of concern. In education, scientific research study is chiefly concern with the variation observed in human behavior, why such variation occur and how it could be controlled or varied desirably. Some of such variations constitute a problem for the process of educating, and it is through the study of the sources of such variation that solution to such problems could be found.

The finding of educational research as a means of searching for the truth about human behavior is valid to the extent that we can maximize the strength of the influencing or our independent variable; control influence of extraneous variables and reduce the amount of random or unintended error committed during the process of research. Such error emanates from the way we sample and the way we measure our research variables. The later involves instrumentation and data collection processes. The influence of extraneous variable constitutes an error in research, since such error is predictable, it is said to be systematic, while the influence of unintended error is random and unpredictable. Kerlinger's MAXMINCON principle (Kerlinger & Lee, 2000) which is designed to enhance validity in research finding can also be adapted for assessment (see Table 1).

Both research and assessment are involved with variables. While assessment defines and quantifies a variable, research tries to determine to what extent and why a variable varies. Hence both are involved with studying a variable by defining, describing and operationalizing it, and by analyzing it to determine the extent to which and why it varies. Scientifically one cannot analyse that which is not measureable. So while assessment through measurement defines and quantifies a variable, research analyse it to determine what makes it a variable, that is, what makes it vary and how and why two or more variables co-vary.

Research Error

Two sources of error are involved in research: error in sampling and error in themeasurement of research variables. Error in the measurement of research variable involves error in the construction of the measurement instrument, and error in the process of using such instrument in the actual process of data collection.

In research, random error is the unintended error that is an inevitable part of the sampling process. It is fluctuating and unpredictable, and it means out to zero. In other words, it does not affect the size but the distribution of sample (see Table 1). If several, say 100 samples of the same size, say 30 (n = 30), are taken from the same population, we will have 100 means of what we are measuring from these 100 samples. These are expected to be the same as they represent parameter of the same population. In actual practice we always take one sample only to represent the population. So how much error do we commit by doing this? The expectation is that since each of our 100 samples is meant to be representative of the same population, the sample mean of what we are measuring for the 100 samples should be the same. But in actual practice this is never the case. The 100 means differ to the extent that we have committed unpurposeful error in sampling. If we found the grand mean of the 100 means, we will get a good estimate of the actual population mean in whatever we are measuring. The difference between this grand mean and the mean of each of our sample is indicative of the error we commit during each sampling. The standard deviation of such differences gives us the sampling error of the mean.

Since we always sample only once during a research study, we use the standard deviation of our one-time measure to estimate the sampling error called the sampling error of the mean with the following formula:

$$S_{mean} = \frac{S}{\sqrt{n}} \qquad 1$$

Where S is the standard deviation of the one-time measure of say 30 different research subjects (n = 30). Let's say in our measurement of our variable for these 30 subjects we have a standard deviation of 4.66; then our standard error of the mean which represents the size of our random error would be:

$$S_{mean} = \frac{4.660}{\sqrt{30}}$$

$$S_{mean} = \frac{4.660}{5.477}$$

$$= \quad 0.851$$

Hence the size of standard error of the mean representing random error due to sampling is 0.851. This indicates how our means, which are expected to be the same, would vary given different samples from the same population.

Note that the size of this error depends on two things: the size of the standard deviation and the size of our sample (n). The smaller our sample standard deviation, the smaller will be the size of our random error; the larger our sample size the smaller would be the size of our sampling error or random error due to sampling. A large simple random sample from a normally distributed population has been determined to give the least possible value of standard error of the mean.

For the sensitivity of our research study, it is important that S_{mean} be kept as small as possible, because it provides the benchmark based on which our null hypothesis is tested. It is that to which we compare the variability brought about our independent or manipulated variable on our dependent or problem variable to, to determine whether such effect is significant. So when we say 'over and above that due to error', it is this error that is referred to.

Hence the smaller S_{mean} the higher the probability our rejecting a false null hypothesis and the hence the higher the power of our statistical testing.

Being an inferential process error is necessary to support probabilistic estimation. The estimated value of random error provides the benchmark for determining the significance or not of the influence of an independent variable (IV). The influence of IV is significant to the extent that variability of the problem or dependent variable due to it is over and above that due to random error. How many times does the variance due to IV over and above that due to random error?

Table 1

Applying Kerlinger's MAXIMINCON Principle to Improve Assessment and Research

#	Aspects of MAXMINCON	Research	Assessment
1.	Maximizing the systematic or experimental desirable variance (MAX)	Treatment conditions should be pulled-apart, made to differ or maximally differentiated among experimental levels as much as possible. That which makes experimental groups different should be sharp, focused and should ensure as much mutual exclusion as possible. Even in non-experimental settings levels of the independent variable should be differentiated as much as possible.	Use measurement instruments with highly discriminating item. Item response theory analysis should be used to select items, because it could detect discriminating items at every point of the ability level.
2.	Minimizing the random error variance (MIN)	Random error in educational research are measurement- and sampling-related, hence standard error of the mean and standard error of measurement are two components of research error whose effect serves as the denominator when determining significance ratio during statistical testing. Measurement involves the quality of the instrument with regards to its validity & reliability, and the quality of data collection procedure. This error is minimized by developing highly reliable instrument, sampling scientifically, using large sample size and a rigorous data collecting technique.	Reduce error of measurement. Use assessment instruments with high reliability. Carry out scientific sampling from a well-defined content and behaviour domains. Large sample size in subject (or participants) and item samples enhances validity in research and measurement.

3.	Controlling the systematic undesirable variance (CON)	Systematically control, isolate, rule out or eliminate the influences of variables extraneous to the relationship under study through statistical or experimental methods especially randomization. Invalid operationalization/measurement of research variables, including bias in data collection process; sampling bias are rich sources of systematic error in research as they tend to make research findings not what it would have been if another definition or sample from the same population is used.	Pre-test instrument and check for and eliminate factors with systematic extraneous influences, like item bias, on performance. Application of the principles of item response theory in test construction and analysis. With its assumptions of unidimensionality and local independence, IRT provides for the control of systematic extraneous variance.
4.	Internal validity	The degree to which that which is being manipulated or the independent variable is that which brings about the variability in the dependent or problem variable claimed for the independent variable. Random assignment – randomization to control internal invalidity.	The degree to which the result of the measurement directly reflects the level which the observed performance results from only the confrontational interaction between the ability under measurement alone and the cognitive (could be affective or psychomotor too) demand of the item. That is, the ability under measurement alone is that which is being measured The internal validity of a measure is asking the question: to what extent are we sure that the score generated from measuring the learner's ability accrues only from the influence of the ability being assessed? That is, how confident are we about the 'causal' relationship between that which is being measured and the score that results from the measurement?

5.	External Validity	The degree to which the findings of the study reflect the truth and are hence are replicable or generalizable to the population, setting and treatment. Select representative sample through simple random sampling.	The degree to which the result of measuring a variable or an ability with the instrument will correlate with the measure of the same ability using several other instruments designed to measure the same thing. Other than the set of tasks (items), conditions, examiners, etc., used to measure the learner, how confident are we that if we used another set of tasks designed to assess the same ability or behaviour, under a different condition and another set of examiners, etc., we would arrive at the same score? This implies ability to generalize, based on the observed score, to the ability of the testees, or the extent to which the observed score could be used as a true representative of the ability under measurement.
6.	Sources of Error	**Sampling**– small sample size and non-scientific sampling leads to sampling bias; that is, systematic sampling error. Poor return rate upsets aims of any scientific sampling plan. Sampling error is implied in the difference between population parameter and sample estimate of such parameters. It is inevitable, but it should be ensured that it is only due to random and not systematic factors. Scientific sampling and large sample size minimize random error and controls for systematic sampling error.	**Instrument Construction -** Non-scientific operationalization of variables/constructs, including poor domain definition and non-scientific sampling of indicators from the domain. Hence lack of representativeness of indicators used to construct instrument. Impinges on the validity of resulting scores and hence on generalizability of score to measured ability/behaviour

7.	Sources of error (cont'd)	**Measurement** – Poor quality and length of instrument. Non-scientific operationalization of variables /constructs, including poor domain definition and non-scientific sampling of indicators from the domain. These lead to lack of representativeness of indicators used to construct instrument. Impinges on the validity of resulting scores and on generalizability of score to measured behaviour.	**Data Collection Procedure**. Poor administration procedures. Invalid and incomplete responses from participants. Poor response rate upsets aims of scientific sampling.
8.	Sources of error (Cont'd)	Data Collection Procedure – Poor administration procedures. Invalid and incomplete responses from participants. Poor response rate upsets aims of scientific sampling Inaccurate handling of data – Inaccurate scoring and coding of data. Poor reliability and validity of instrument and hence invalid operationalization of variables. Using invalid statistical analysis techniques.	Inaccurate handling of data –Inaccurate scoring and coding of data. Poor reliability and validity of instrument and hence invalid operationalization of variables. Using invalid statistical analysis techniques
9.	Type of Validity of most concern	Scientific research is a theory-based exercise which involves the conceptualization and operationalization of several constructs in its process path. So educational research is more closely involved with construct validity than with the other types of validities to which it is also related.	The process of developing a valid measurement instrument is tedious and scientific. Depending on whether it is used as a tool for assessing achievement or ability. Most abilities or behaviour are latent and need to be conceptualized and operationalized also. But for achievement test, measurement as a tool of assessment is mostly concerned more with content validity, as the contents of two domains must be elaborately defined sampled and used to operationalized the

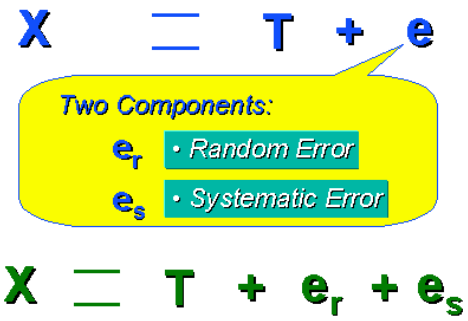
10.	Reliability	To what extent does a finding in one trial or one study persist across several similar trials or studies? Persistence of a finding across trials of the same study is indicative of internal validity. That is, the influence of the independent variable on the dependent variable is persistently the same across repeated similar studies.	To what extent does performance in a test repeated during similar or repeated testing with the same, alternate or parallel instrument? That is, to what extent does the result of testing devoid of random error – error that cannot be consciously repeated? How well does an instrument consistently measure what it is measuring? Reliability reflects the precision of measurement.
11.	Measurement Error		

Figure 2 Illustration of Measurement Error
(Source: Trochim, W. M. K. (2006))

12.

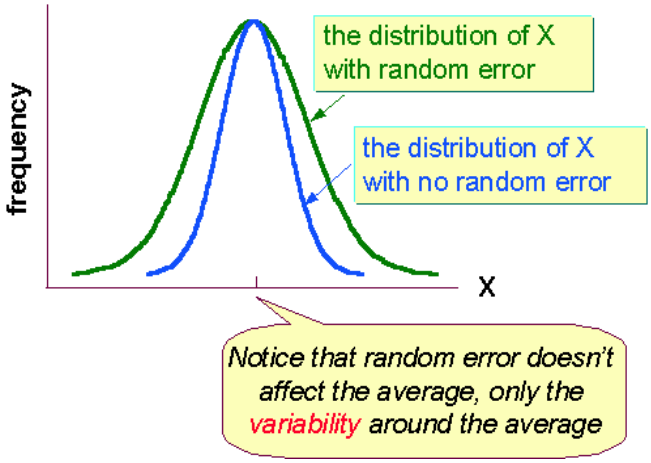


Figure 3 Illustration of Effect of Random Error on Score Distribution
(Source: Trochim, W. M. K. (2006))

13.

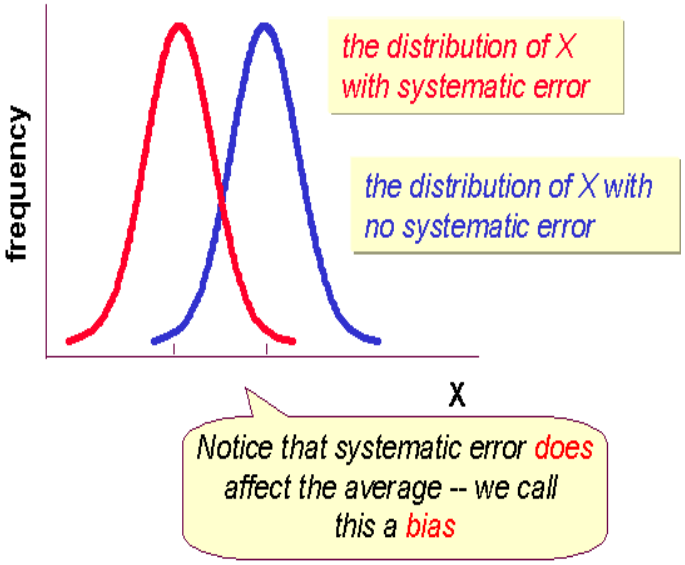


Figure 4 Illustration of the Effect of Systematic Error on Score Distribution
(Source: Trochim, W. M. K. (2006))

Being an inferential process error is necessary to support probabilistic estimation. The estimated value of random error provides the benchmark for determining the significance or not of the influence of an independent variable (IV). The influence of IV is significant to the extent that variability of the problem or dependent variable due to it is over and above that due to random error. How many times does the variance due to IV over and above that due to random error?

Research, Assessment and Theory

Both research and measurement are theory-based scientific exercises. Quantitative research consciously or unconsciously is a theory-validation endeavor (see Figure 5). It is founded on theory whose consequences provide the "tentacle of knowledge" to be tested and if sustained confirms the theory for which it is a tentacle or otherwise refutes it. Thus theory is built, validated, grown or trimmed to size through empirical research. Essential to the validation of any scientific theory is data, quantitative empirical data, which is mainly generated through assessment, especially measurement. Scientific or empirical research cannot

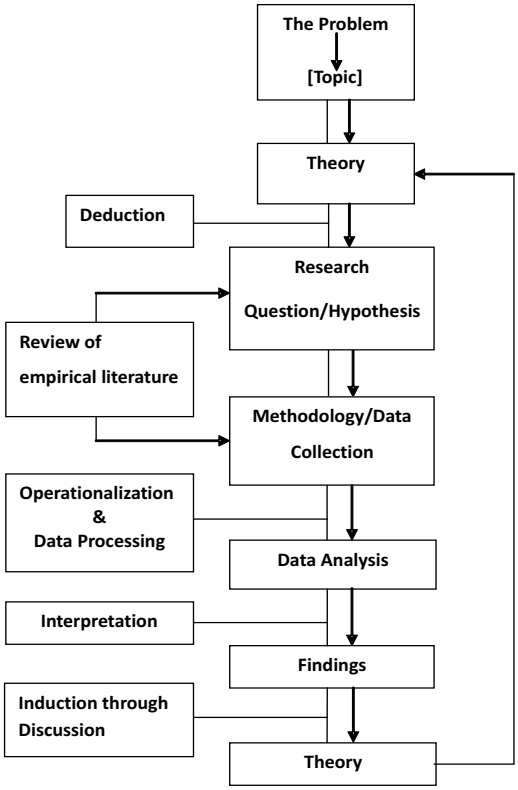


Figure 5. The structure of quantitative research process

proceed without theory and without measurement. Besides that research essentially calls on measurement, measurement is based on theories that are developed and validated through the research process.

Theory is developed through research which also goes to validate it while measurement is based on theory built through research and serves to seek for quantitative inputs into and operationalizes its empirical processes. Both measurement and research are large-number happy concepts (Nunnally, 1978). Parameter estimates based on a sample either of subjects or of items used to measure a behavior tend towards the population value as the sample size or number of items used in the measurement tends to the population size. For a better estimate of achievement, forms of assessment other than measurement should also be used. Using several assessment tools and repeated assessment provide opportunity for ability to show up all dimensions of its potency. Hence continuous assessment is often recommended (Nenty, 1991).

A more valid estimate of ability is assured if it is given several opportunities to confrontationally interact with many appropriate and varied tasks that can provoke such ability to action. Hence, for example, longer tests are generally more reliable and even more valid than shorter tests.

Conclusion

Underlying research and measurement is the yearning for validity or scientific truth. Both are inferential processes of seeking for truth about latent characteristics or properties of individuals as well as populations. For psychological measurement, we seek to estimate the amount of an ability possessed by individuals by challenging them with tasks carefully developed to provoke such ability. In research, we seek to estimate a characteristic of a given population based on what we observe from sample drawn scientifically from it. Such estimations necessarily involve error, which is either systematic or random. These, depending on their sizes, mar the chances of what we observe being a good estimate of the true value of what we were estimating. To the scientist, there is always the need to maximize the variance due to the ability or parameter one was trying to estimate; control the extraneous variance due to systematic sources of error and minimize the ever-present variance due to random error. In assessment and research, to get at that which is concrete from that which is abstract, we need the guidance of a theory, hence both research and assessment, especially measurement, are theory-driven.

Reference

Biesheuvel, S. (1974, April).The use of ability tests in developing countries: Some comments on Ord's monograph. *Psychologia Africana, 15*(2), 119-126

Brubacher, J.S. (1939). *Modern philosophies of education*. New York: McGraw-Hill Book Company, Inc.

Dhaliwal, I., & Tulloch, C.(n.d.). From research policy: Using evidence from impact evaluations to inform development policy. Retrieved from: http://www.povertyactionlab.org/publication/research-policy

Kerlinger, F. N., & Lee, H. B. (2000).*Foundations of behavioral research*. Singapore: Thomas Learning Inc.

Kerlinger, F.N. (1986). *Foundations of behavioral research* (3rd ed.). New York: Holt, Rinehart and Winston, Inc.

Lord Kelvin (1883-05-03). *Electrical units of measurement. PLA, vol. 1, Retrieved from:* http://zapatopi.net/kelvin/quotes/

Nenty, H. J.,& Lusweti, S. L. (2014). Assessment for learning (AfL): Implications for the achievement of the goals of basic education in Africa. *African Journal of Theory and Practice of Educational Assessment* (EARNiA Journal), 1, 34-51

Nenty H. J. (2000).Some factors that influence students' pattern of responses to mathematics examination items. *BOLESWA Journal of Educational Research, 17*, 47 – 58.

Nenty, H. J. (1997a, July 26-August 1).*Links among education, research, educational research, and quality of life.*(Invited lead Paper) Seventh BOLESWA International Symposium at the University of Swaziland, Swaziland, on July 28 - August 1, 1997. Retrieved from: http://boleswa97.tripod.com/nenty_link.htm

Nenty, H. J. (1997b, July 26-August 1).*Assessment as a means of enhancing improved quality of life through education*. Seventh BOLESWA International Symposium/Conference at the University of Swaziland, Swaziland, on July 28 - August 1, 1997. Retrieved from: http://boleswa97.tripod.com/nenty_assesment.htm

Nenty, H. J. (1991/92).The basis of education science. *Eduscope, 5*, 8 – 11

Nenty, H. J. (1991, August 26) *Introduction, concepts and practice of continuous assessment.* Seminar on the New Education Policy and School Management in Ikom, Nigeria.

Nunnally, J. C. (1978). Psychometric theory. New York: McGraw-Hill Book Company.

Pohjola, M. V., & Tuomisto, J. T. (2011).Openness in participation, assessment, and policy making upon issues of environment and environmental health: a review of literature and recent project results. *Environ Health, 10*, 58.

Trochim, W. M. K. (2006). *Measurement error*. Retrieved from: http://www.socialresearchmethods.net/kb/measerr.php/