CLASSICAL TEST THEORY'S AND ITEM RESPONSE THEORY'S PERCEPTIONS OF THE VALIDITY OF TEST SCORES

H. Johnson Nenty, PhD.
Educational Foundations, University of Botswana, Botswana

Abstract

In testing, performance is that which results from the confrontational interaction between the amount of ability under measurement possessed by a testee and the amount of trait demanded by the tasks developed to evoke such ability during testing. The interpretation of such performance given the prevailing values and standards gives achievement. Hence, test scores have meaning or test scores are valid to the extent that it was only the ability under measurement that sustained the responses to each item in the test through which the scores were generated. While classical test theory thrives on reliability, that is, how well a test repeatedly gives same score; item response theory sets to ensure that the score generated through a test represents the truth. That is, CTT is tilted towards satisfying the dictates of reliability, the repeatedness of the same test score across time, form, and parts of the same test, IRT is tilted towards ensuring a valid test, that is, ensuring that it is only what the test was designed to measure that underlies responses to its items. In other words, reliability is closer to the heart of CTT than that of IRT hence CTT is oftentimes referred to as reliability theory. For CTT, validity is faintly defined because there is no attempt to ensure that what underlies responses to items in a test is only that which the test was designed to measure, whereas for IRT the assumption of unidimensionality implies its recognition and step taken to ensure that items that fit any of its models measure one and only one ability.

Keywords: Classical test theory; item response theory; reliability; validity; unidimensionality.

Introduction

The idea of measuring human behaviour, a concept that cannot be seen, felt, smelt, or touched is intriguing, and to some people, it is impossible or a big joke. Even the attempt to measure the physical, that is, that which can be seen, felt, touched, etc. is never without error. But if education, which is often defined as 'everything done to ensure and maximize desirable changes in human behavior,' is to be accomplished, then something must be done to document how much of such changes our efforts as educators, have brought about and factors that can maximize such changes. To Sir 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. . . ." There is yet no art through which we can just look at the learners on the face or open up and inspect their brains to determine what they know and how much they know it, or what they can do, and how much of it they can do. Hence educational assessment, which is, the art and science of quantifying or qualifying the cognitive, affective and psychomotor behaviour of learners, is inevitable.

The practices of such art and science are guided by some principles and understanding often termed measurement models or theories. Until about the middle of the last century the practice of measuring human behavior was based only on a model often popularly called classical test theory (CTT). Of more recent development is the item response theory (IRT).

Common to both models is the belief that each learner has and carries around some amount or level of behaviour which is often called ability or trait (Warm, 1978). That is, every given ability or trait is a psychological [cognitive, affective (emotional) or psychomotor] property of an individual and every person possesses each of these to some degree. One person cannot have two values of the same ability at the same time. The same is true of physical traits; but these we can observe and measure directly. That is, we can measure these physical traits with instruments which during the measurement process have some kind of direct physical contact with the trait in question. For example, when we want to measure a physical trait called height, we make the carrier of such trait to have a close and direct contact with a rule calibrated to indicate the amount of that trait possessed by the body, and then read the height as the point on the rule that corresponds to the amount of height possessed by the body.

This is a direct measurement. We cannot observe psychological traits, let alone to have an instrument that can have direct contact with them. They are hidden or latent traits and hence can only be measured indirectly, that is, by prompting their reaction to well-prepared tasks that can only be overcome by the exhibition of the latent trait in question. Such prompting or confrontation calls or provokes to action the amount of that trait or behaviour possessed latently by the person or body. In other words, while physical traits are measured directly, psychological traits are measured indirectly or inferentially because their size can only be inferred from the results of their interaction with tasks carefully designed to elicit their presence and potency. Given this scenario, we need a psychological or theoretical touch-light to guide our steps while looking for that which we cannot see, but can only infer through the potency of its action, that is, that which we cannot see but can only observe the consequences of its potency. Hence, a test theory or model provides a general framework linking observable variables, such as test and item scores to latent or unobservable variables, such as true and ability scores (Lord, 1980).

The unseen or latent cognitive, affective and psychomotor traits are properties of every individual and we carry them along wherever we go. Each can be elicited or provoked into action by challenging it with relevant tasks. Such tasks act as stimuli to evoke the ability implied by the trait. This rises to the occasion every time such or similar task is encountered.

Galileo's popular advice is that we should *"Measure what is measurable, and make measurable what is not so."* is exemplified in the measurement *of* speed of the unseen-able wind (W_s). This is measured through obstructing its current by placing an obstacle called the wind-vane (V_r) on its path. In other words, we try to gauge the speed of the wind (W_s) which we can only feel but cannot see, by tasking the strength of its current with the resistance (V_r) of an obstacle called the wind-vane. The higher the speed of the wind, the faster the vane will rotate. The probability of turning the wind-vane and the speed of its rotation depend on the difference between the speed of the wind and the resistance of the vane. Given these settings, the initial concern of the paper is to come up with some fluid consideration of basic concepts in measurement followed by a discussion of equally fluid views of how CTT and IRT perceive, define and operationalize validity.

Education, as seen earlier, is the process of ensuring and maximizing desirable changes in learners' cognitive, affective and psychomotor behaviour. In a formal setting, such behaviour are as stipulated in the curriculum, and detailed in the different syllabi provided to schools. At the implementation level, they are operationalized by behavioural objectives in the different lesson plans developed to guide teaching in the different subject areas. Each of the objectives brings together the subject matter contents and the type and level of human behavior (cognitive, affective or psychomotor) to be changed. For example, an objective like "at the end of the lesson,

pupil will be able to sort and classify into four self-created categories twenty common observations in their environment" combines subject matter – 'environmental science'; with cognitive behavior – 'ability to sort and classify'. A behavioral analysis of this cognitive behavior given Bloom taxonomy may identify: knowledge, comprehension, application, analysis, synthesis and evaluation skills. A syllabus is valid to the extent it is able to operationalize all contents as well as the behavioural objectives implied in curriculum. A valid syllabus provides a foundation for valid implementation of the curriculum through teaching. Similarly, teaching is effective to the extent that it involves and covers all aspects of the syllabus contents as well as all cognitive behaviour as provided for in the curriculum. A guide to item development to measure the extent to which the objectives in the curriculum are met is developed as a table of specification. Like the objectives, item developed must also, bring the subject matter content – environment science - and the cognitive behavior, like analysis, together. Changes in behavior are defined and monitored by assessment.

Items are developed as challenges or tasks which when encountered by the testees will provoke or call into action the appropriate trait or ability under measurement latent in the testee. Appropriate here implies that the item must be craftily developed to target or call only on the ability under measurement. According to Nenty (1985, 2015), a test as a measurement instrument is a set of questions, tasks, statements, etc. (commonly referred to as items) which are developed, tested out, selected, calibrated and arranged in a systematic way as a series of challenges or task to which when an individual is exposed, under controlled condition, elicit the level or amount of the attribute under measurement which the person possesses. One person cannot have more than one value of the same ability at the same time - the time of measurement. The only thing that varies are the debris or extraneous factors surrounding this ability during measurement.

For an achievement test, an item is a translation of subject matter content into tasks, questions or statements which when read or confronted by a testee elicits or provokes into action the level of intended cognitive trait or behaviour specified in the lesson objectives possessed by him/her. Hence, each testee's response to each of these tasks indicates how much of the indicated latent behaviour he/she has acquired, given the subject matter, during the lesson or course. Each item is a cognitive, affective or psychomotor task which has some level of ability to provoke or elicit the unseen behavior or trait under measurement possessed by a testee. Thus an item brings together two things: (i) an aspect of the subject matter content, and (ii) cognitive behaviour as indicated in the lesson, course or instructional objectives (Nenty, 1985).

Given the ability under measurement only, the more of this ability an item could provoke before it could be overcome, the more demanding the item is in the measurement of that ability, trait or behavior. In other words, items vary in the maximum level of a particular trait they demand before they could be overcome, that is, some items are more trait-demanding than others. The trait-demand level of an item is the property of the item which is invariant across anybody taking on the item. At the time of interaction with a person's ability, an item cannot have more than one value of cognitive demand at the same time - the time of measurement. The only thing that varies is the debris or extraneous factors surrounding this cognitive demand.

Similarly, every testee carries around some level of any given trait (Warm, 1978), and some testee has more ability- or are richer in a particular trait than others. Such ability or trait level is the property of the person and is invariant across several tasks demanding that ability before it could be overcome. The probability of overcoming an item that demands the ability under measurement possessed by the testee is dependent on the difference between the amount of that ability possessed by the testee and that demanded by an item before it could be overcome.

During testing the level of ability demanded by each item remains constant across all examinees, while the ability-value of each testee also remains constant across all items. This is the basic idea of invariance, that is, the ability parameter of a testee for a given measurement remains constant irrespective of the items used in the measurement. Similarly, the cognitive-demand or difficulty of a task or an item remains constant across the level of ability possessed by all examinees that take-on the item. In other words, item and person parameters are invariant.

Ability and Achievement

Performance is that which results from ability-by-task confrontational interaction (see Figure 1). There are factors other than ability under measurement that comes in to play during the process of such interaction. Ability or trait score can only be validly estimated if the usual requirement of 'under controlled (physical and psychological) conditions" is ensured during testing so that it is only the trait under measurement that is underlying the responses during this interaction. Such conditions, if not controlled, spill extraneous influence all over the testing atmosphere, and inhibit the achievement of valid scores from the ability-by-task interaction.

Using a magnifying glass to look at what goes on during psychological measurement it could be seen that factors that are extrinsic to the measurement purpose come into play during the ability-by-task interaction. The result of such interaction could only be used to estimate ability validly to the extent that it was that ability alone that influenced what went on during the process of the interaction. Achievement is the interpretation of performance given the prevailing values and standard. For example, scoring 60% in a mathematics examination is a mathematics performance, based on this score, the testee can be assigned a passing or failing status in mathematics given the test, or can be awarded a grade of D or C or B depending on standard used to translate score into grade. This is the testee's achievement in mathematics. Since performance also depends on factors other than ability, achievement tends to vary with these other factors that influence the interaction process, hence with the same ability but different extrinsic factors, different performance and hence achievement may be observed. Hence, two testees, with the same level of ability under measurement, but who are differentially influenced by different extrinsic factors or at different level of the extrinsic factors, could be observed to score differently in a CTT-based testing. Hence, equality in CTT scores does not necessary implies equality in the possession of the ability under measurement, and *vice versa.* Earning the same or equal scores generated through CTT-based measurement does not necessarily translate into the possession of equal ability. Similarly, two items with the same level of cognitive demand but different levels of demand of extrinsic ability will tend to draw different responses from teetees even those with the same level of ability. For example, while the weight of a piece of stone is an inherent property of the stone, the stone tends to be heavier to a child than to a grown up person. In that case, age is an extraneous factor. Hence, the CTT estimation of person ability depends on the test he/she is taking while the estimation of item difficulty depends on the group of testees to whom the item is administered. Given a measure of the same ability, a teacher can decide to make his/her students look smart by giving them a test made up of 'easy' or cognitively less demanding items, and can as well decide to make them look 'dumb' by giving them a test made up of 'difficult' or cognitively highly demanding items. Wright (1967) talks of it as measuring with an elastic meter rule.

The Classical Test Theory (CTT) (synonymous to True Score Theory)

The classical test theory, or commonly termed as true score theory, postulates that the score we observe (X_o) for a testee on a test or raw score, is a component of two scores: the true score (X_∞) and error score (X_E) and they are related through the formula;

$$\mathbf{X_o} = \mathbf{X_\infty} + \mathbf{Xe} \quad (1)$$

The bigger the error score, X_e, the less the observed score estimates the true score. Error score X_e is the uncontrollable error that results from fluctuating or random influences related to the testing conditions, the emotional and health state of the individual testees, and the quality of the test items which unsystematically influences testees' responses during testing. It is random and hence results from the operation of chance. Being random and chance-driven, it causes measures to unpurposefully fluctuate around the mean of what is being measured across testees and hence does not affect the overall mean of the group, but the variation of test scores. In a measure, the sum of all random error is zero; hence the mean of random error scores is zero. Hence, random error does not change the value of observed score mean, but its distribution.

Equation 1 as stated is unsolvable because there are two unknowns. An attempt to solve this call for some basic assumptions: (i) X_∞ and X_e are uncorrelated, that is, are independent in that case, their variances are additive; (ii) the average of X_e in a population of testees is zero; and, (iii) X_e from parallel tests are uncorrelated. With these assumptions, it could be shown that X_∞ to be the expected observed score across parallel forms (Lord & Novick, 1968). Parallel test forms are test developed based on the same table of specification and across which the testees have the same true and error scores. Lord and Novick present an equivalent of $\boldsymbol{X_{int} + X_{ext}}$ as "redefined true score." (pp. 43-44)

Reliability as an estimate of the consistency of performance of the same instrument, indicates how well a test relates to or correlates with itself across time, forms of the same test, items in the same test or across sections or parts of the same test. It is akin to establishing the consistency of a story told exactly repeatedly even though it might be a lie, whereas validity is concerned with the extent to which the story represents the truth. In estimating reliability, we do not need an external criterion to which we compare our measures. To reliability, that which makes test scores not to repeat itself if the test is taken more than once with the same instrument is random error. Hence random error causes score from the same test not to be consistent, repeatable, trustworthy, and stable across time, forms, parts of the test and test items. For a test, the bigger the random error is, the less reliable the test.

Since reliability does not take into consideration any other sources of error except random error, it is very close to the heart of CTT which also does not emphasize any source of error other than random error. Hence, the classical formula relating observed score to true and error scores is given by Equation 1. Given the assumption that X_e and X_∞ are not related then the variances related similarly thus:

$$V_o = V_\infty + V_e \quad (2)$$

Reliability, as the degree to which the observed score is free of random error is therefore defined as:

$$r_{tt} = \frac{V_\infty}{V_o} \quad (3)$$

$$= \frac{V_o - V_e}{V_o}$$

$$= 1 - \frac{V_e}{V_o} \quad (4)$$

If all items in a test are designed to measure the same thing, whatever that is, the items are seen as replicates, then any within-item variation is error. This is the variation above and beyond that due to the true ability, and this for CTT is random error. Hence, the sum of such variations estimates V_e,. A close inspection of Equation 2 shows that the larger V_o and the smaller V_e the higher the test reliability. Coefficient alpha as an index of reliability is based on this logic. That is:

$$\alpha = \frac{k}{k+1}\left[1 - \frac{\sum V_{ei}}{V_o}\right] \quad (5)$$

Where α = coefficient alpha
K = number of items.
The ratio, $\frac{k}{k+1}$*, is introduced to constraint alpha to range between 0 and 1;*
$\sum V_{ei}$ *is the sum of the variations for all the items, whereas* V_o *is the variance of the observed test score.*
Reliability is the degree to which an instrument is consistent at measuring what it is measuring. It might be measuring influence of biasing or extraneous factors; it is reliable as long as it is measuring this consistently.

Systematic Error Score and Validity

CTT does not provide for, nor take into consideration non-random error. Since this is systematic, it constitutes a component of the true score variance which is also systematic and hence loads this variance predictably. While random error affects the precision of measurement, systematic error affects the accuracy of scores. In other words, a test can be precisely measuring that which it is measuring (precision) without measuring that which it was designed to measure (accuracy). Hence, while observations must be precise in order to be accurate not all precise observations are accurate. In other words, precision is a necessary but not a sufficient condition for accuracy, or, reliability is a necessary but not a sufficient condition for validity.

Systematic error is predictable error emanating from the influence of biasing factors such as language in a mathematics test. Over and above the ability under measurement, sources of systematic error 'cause' the testee to score more or less than he/she would have scored if its influence was not there. It is an unwanted source of influence during the ability-by-task interaction. Item response theory (IRT) works validly only if sources of such extraneous influences are controlled to a non-significant level. This is a more realistic perspective than that held by CTT, as seen earlier (see Equation 2), and calls for the adjustment of the variance involved in the classical test formula to:

$$V_o = V_{ability} + V_{extraneous} + V_e \quad (6)$$

And given this presentations, validity is:

$$Val. = \frac{V_{ability}}{V_o} \quad (7)$$

Because it is only concerned with the contribution of the variance due to that which the test was designed to measure ($V_{ability}$) to the observed score variance (V_o). Comparing Equation 3 and 7, reliability of a measure is always bigger than its validity. Hence IRT which insists on unidimensional test targets the contribution to the observed score variance (V_o) by variance due to the ability under measurement ($V_{ability}$) only.

Systematic error score, on the other hand, results from the influence of some extrinsic and predictable sources of error due to biasing factor or factors that influence the ability-by-task interaction condition (see Figure 1). The influence of systematic sources of error, like biasing factors, for example, illustrations and examples used in the items, the level of language sophistication demanded in a non-language test, and demands on the testees other than those related to what the test was actually designed to measure. The error that emanate from these systematic sources forms a part of X_∞. Hence, according to Harvill (1991), "The term *true score* is a bit misleading because any *systematic* error such as examinee's reading ability or test-taking skills (test-wiseness) is considered part of the true unchanging portion of an examinee's observed score" (p. 33). Also, "as used, true score is not an ultimate fact in the book of the recording angel. Rather, it is the score resulting from systematic factors one chooses to aggregate,

including any systematic biasing factors that may produce systematic incorrectness in the score" (Stanley, 1971, p. 361). Hence, the classical test theory misrepresents what actually goes on during measurement by implying that true score, X_∞ represents what one was trying to measure only. No, it represents the influence of all sources of systematic variation on the observed test score.

Based on these understanding, for an item Formula 1 could be modified (Biesheuvel, 1974, Nenty, 2000) to:

$$\boldsymbol{X_o} = \boldsymbol{X_{ability}} + \boldsymbol{X_{ext}} + \boldsymbol{X_e} \quad (8)$$

[*A modeling that differentiates between that which represents the ability under measurement, X_∞, and that which represents all other factors X_{ext} that have systematic influence on observed score, X_o during the measurement process (see Figure 1). To control for this sources of error, IRT assumes unidimensional measures. If variance due to X_{ext} is significant, that is if V_{ext} is significant, that is over and above V_E, then IRT is short-chained and its desirable properties cannot be realized. X_{ext} includes the score made under the influence of systematic extraneous abilities like language in a mathematics test, guessing, bias, etc*.]

Whereas, validity could be said to be the extent or probability to which an instrument is consistent at measuring what it is intended to measure. In other words it is the extent to which scores derived from our using our instrument in measuring reflect what the instrument was designed to measure. It is the extent to which our instrument is not measuring extraneous or biasing factors but the amount of the trait is was designed to measure possessed by the body. Hence, unlike CCT which is reliability minded, IRT which insist on unidimensional measure, is validity-happy.

For CTT, item statistics include item difficulty (p-value) which, for objective items, is the ratio of the number of testees that got an item correct to the total number of testees that attempted the item. For an essay item, this is the ratio of the mean observed score on an item to the maximum possible score for that item (Neny & Umoinyang, 2000). The ideal p-value is .50. Hence, for CTT the difficulty (p-value) of an item is not an inherent property of the item but depends in the ability of the testees to whom the item is administered. If it is exposed to another set of testees with different ability, it will automatically change its difficulty. What a measurement!

Another important CTT item statistics is the item discrimination (d-value). This is how well an item differentiates between those who have a good level of knowledge of what the item is measuring from those who have poor level of knowledge of it. One way of estimating this index is to find the difference between the difficulty of the item for the upper third-scoring group in the entire test and that of the difficulty of the same item for the lowest third-scoring group in the entire test (Nenty, 1985). Another estimate of the item discrimination is the item-test correlation value. That is the value that results from point bi-serial correlation analysis between the item scores and test scores of all testees. Though items with d-values of .40 and above are said to be good, the higher a d-value the better the item. Another item statistics is the guessing index. This, for CTT is the reciprocal of the number of options for an item. CTT item discrimination index suffers from the same handicap as its item difficulty index.

For CTT person statistics is the total score, which in a multiple choice item, is the number of items in a test that a testee was able to choose the correct option. CTT uses this as an estimate of a testee's ability. In that case, it equates achievement selfishly to ability. Performance is ability driven, that is, it is underlay by ability. It results from the confrontational interaction between ability and an appropriate related task. While CTT defines validity in the context of

'true score' the sources of systematic variation in observed score, IRT defines it in the context of 'trait score' While the absence of random error enhances reliability, the absence non-random error in measurement enhances validity. Non-random error affects, predictably, the values of the observed score, and hence its mean.

CTT is a test-based theory. Hence, it is very sensitive to test parameters, especially test reliability (r_{xx}) .This is the consistency with which a test measures what it is measuring. It represents the precision of measurement, how well measurement result, that is X_o is free from random error. Given the CTT assumption that X_∞ and X_e do not correlate, observed score variance could be said to be equal to the sum of true and error score variances.

A more important index of test quality is the validity index. At the testees' level, validity is seen as the degree to which their scores represent the level to which they possess that which the test was designed to measure. Generally, it is an indication of the level to which a test is measuring what it was designed or used to measure. That is, the level to which a test score reflects or is telling the truth about what the test was used to measure possessed by the testee. If what the test was designed to measure is known then this would be the variance of the observed test score divided by the variance of that which the test was designed to measure. This could be considered as the proportion of the trait or ability score variance to the observed score variance. Kerlinger and Lee (2000) see it as the proportion of common factor variance to total variance of a measure. In the same consideration, item validity results from dividing the observed score variance by the variance of the ability or trait which the test was designed to measure.

The common factor variance represents the trait variance, that is, the variance of the trait which all the items in the test were designed to measure, and hence the variance which is common to all of them. Besides this variance, the observed test score variance also contains the systematic variance from extraneous variables which influences the results of the person-by-item interaction besides the trait variance. Such extraneous variables include biasing factors like the language of the test, every other variables, besides the trait that has systematic influence on testees' responses to the items. While the trait variance could be said to be intrinsic (V_{int}), the systematic error variance could be said to be extrinsic (V_{ext}) to the interest of the measurement (Biesheuvel, 1974; Nenty, 2000). Assuming an independent relationship between $\boldsymbol{X_{int}}$ and $\boldsymbol{X_{ext}}$, Formula 2 could be rewritten as:

$$\boldsymbol{V_o = V_{int} + V_{ext} +} \text{ a Ph. D student at} \boldsymbol{V_e} \quad (9)$$

Given Formula 5, validity is represented by (Kerlinger & Lee, 2000):

$$\boldsymbol{Val} = \frac{\boldsymbol{V_{int}}}{\boldsymbol{V_o}} \quad (10)$$

Comparing Formula 6 to 4, given Formula 5, "systematic biasing factors such as test-wiseness do not affect test reliability but certainly can negatively affect test validity" (Harvill, 1991). Hence, the value of test validity is always lower than that of reliability of the same test.

Establishing evidence of validity:

1. At the qualitative or non-technical level, given an expert's knowledge of what the instrument is designed to measure, to what level does each item looks like, it is measuring that for which the instrument is constructed? Such level can be quantified by having the expert to assign a relative rating. Here, the criterion – what the instrument is designed to measure - is as defined by the expert. The expected rating can be used to establish content validity as explained by Lawshe (1975).

2. How much does the variability of the observed scores arrived at by using your instrument accounted for by one single source which represents that which we designed the instrument to measure?
3. How well do the scores observed using this instrument or this item relate to scores that are rich in representing that which the instrument was designed to measure?
4. To measure a given construct, how well do the scores generated based on your instrument relate to those generated for similar constructs but based on other methods of measuring the same constructs but fail to relate to scores generated using the same measurement methods but measuring different constructs?

Note that, except for the second evidence, in each of these means of establishing validity, scores on an external criterion assumed to be rich at measuring the same construct, trait or ability are involved. Hence, scores on your new instrument are compared to those from another instrument that represents the truth, given what your instrument was designed to measure. To what extent is your instrument telling the truth? For the second evidence, the thrust is to extract that which is common given the variability of all the items constructed to measure the same trait or ability in question (See Venn Figure 1)

CTT to IRT Transition

Technical Consideration of the Relation among Testing, Score, Performance and Achievement

Testing is seen here as the provision of a physically and psychologically conducive environment that maximizes the result of the confrontational interaction between a testee's ability *(θ)* and an item's cognitive demand *(δ)* (See Figure 1). The result of this interaction is a score (X). The interaction between ability A (infected with A_E - testee-related sources of error) with item demand B (infected with B_E - item-related sources of error) gives $X_{(Score)}$, often termed performance. This when interpreted, given the prevailing values and standard (C),

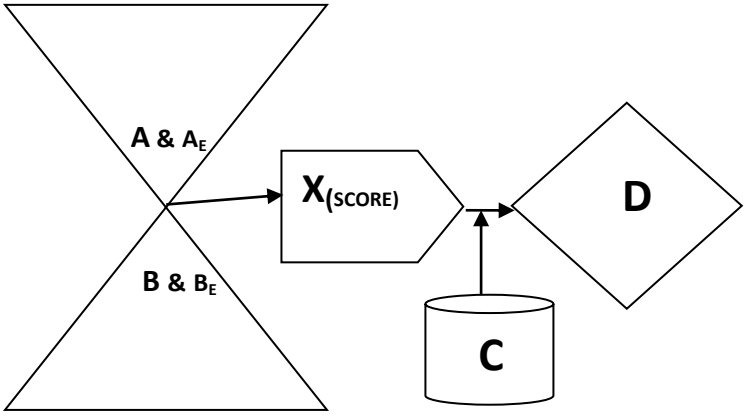

Figure 1: Interpretation of the result from the interaction between testee's ability (θ) and item cognitive demand (δ).

A ⟹ *testee's ability (θ)*

A_E ⟹ *testee's-related sources of error (e.g. sluggishness, cheating, ability to guess, etc.)*

B ⟹ *item demand (δ)*

B_E ⟹ *item's-related sources of error (e.g. language bias, poor expression, proness to guessing, etc.)*

X ⟹ *performance (X)*

C $\Longrightarrow$ *standard, values, etc*
D $\Longrightarrow$ *achievement,*

gives achievement (D). Achievement therefore is like the evaluation of performance, which is interpreting performance in the light of the prevailing values, standard, etc. CTT does not take into consideration both A_E and B_E in its operationalization of test scores and hence of its definition of test validity. To IRT, on the other hand, a test is valid only to the extent A_E and B_E are absent or controlled for during testing.

The Item Response Theory (IRT)

As the name implies, item response theory (IRT) is an item-based theory that models, in probabilistic terms, the result of the interaction between testee's ability and an item cognitive demand during testing. That is, how the relationship between testee's ability and an item's cognitive demand during an ability-by-item confrontational interaction results in (successful or failing) performance (see Figure 1). It sees testing as a situation in which the amount of latent trait under measurement possessed by a testee is challenged by a series of related tasks each demanding some level of ability in order to be overcome. The level of ability is often represented by theta ($\boldsymbol{\theta}$), while that of item demand is often represented by delta ($\boldsymbol{\delta}$)**.** During such interaction, or confrontation, if the testee's $\boldsymbol{\theta}$ is higher than the item's $\boldsymbol{\delta}$ then the probability of the testee overcoming the task is high. If, on the other hand, the item, to be overcome, demands more ability than that possessed by the testee, then the probability of the testee overcoming the item is low. Hence, both $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are defined with the same unit and mapped along the same scale.

The result of the person-by-item interaction depends on the difference between the person's $\boldsymbol{\theta}$ and the item's $\boldsymbol{\delta}$. When $\boldsymbol{\theta}$-$\boldsymbol{\delta}$ is positive, the higher this difference, the higher the probability of a correct response; if negative, the probability of a correct response decreases with increase in the size of such difference. Hence, the difference between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ is the fundamental index underlying the probability [P($\boldsymbol{\theta}$)] of a person with ability $\boldsymbol{\theta}$ answering an item with $\boldsymbol{\delta}$ level of cognitive demand correctly. This is the most important component in all the equations for estimating P($\boldsymbol{\theta}$) by any of the IRT models. Translating Formula 5 to item level given the emphasis of item response theory on item not test behavior, the outcome of a trait-by-item interaction could be presented as:

Observed item variance = Trait variance + systematic extrinsic variance + Error variance

or,

$$\boldsymbol{V_{o(i)} = V_{trait(i)} + V_{ext(i)} + V_{e(i)}} \qquad (11)$$

In other words, the observed item variance is the sum of the variance due to the intrinsic trait the test was designed to measure; the variance due to testee's, item's or test administration's characteristics which extraneously but systematically influences the result of the ability-by-task interaction; and the variance due to random error.

To ensure a highly valid test, the thrust in test construction and administration should be to maximize $\boldsymbol{V_{trait(i)}}$, control $\boldsymbol{V_{ext(i)}}$and minimize $\boldsymbol{V_{e(i)}}$. Hence, Kerlinger's (1968) method for enhancing validity of research finding also holds true for ensuring high validity of measurement. $\boldsymbol{V_{trait(i)}}$ represents the variance of the trait under measurement, to our measurement situation, it is the desirable variance, while $\boldsymbol{V_{ext(i)}}$ represents variance from the extraneous factors. This is systematic but extraneous and hence undesirable to the intention of our measurement. To Messick (1984), it is termed construct-irrelevant variance (CIV) which reduces the validity of test scores and distorts the interpretation we give to such scores. Given these nomenclature Formula 11 holds.

Assumptions of Item Response Theory

The validity of the results of any data analysis, be it statistical or measurement, depends on the extent to which the data analysed meet the assumptions underlying the statistical or measurement model used in such analysis. For example, analysis of variance as a statistical model, makes some assumptions which data have to meet before the result from such analysis could be valid. If data which do not meet the ANOVA model are analysed using the model, the result of such analysis are not valid or do not reflect the truth. Other statistical models, for example, Kruskal-Wallis non-parametric statistical model, could be applied in analyzing such data. The scientific worth of any measurement model as a means of producing valid scores depends on its ability to generate scores representing achievement that is influenced only by one trait or ability. This is why all IRT models assume unidimensionality of measures. According to Hattie (1985)," One of the most crucial and basic assumptions of measurement theory is that a set of items forming an instrument all measure just one thing in common" (p. 139). A good measurement model should provide a guide to the process of constructing and administrating a test in such a way that the results of a trait interaction with the tasks implied by its items depend on, and only on, the values of the trait and that of each item designed to measure the trait. Hence, unidimensionality assumption is fundamental for a valid operationalization of all IRT models applied to data from dichotomously scored achievement items. If a test is truly unidimensional, then the variance common to all the items represents$\boldsymbol{V_{trait(i)}}$. So to the extent that a test is unidimensional to that extent does $V_{trait(i)}$ dominate $V_{o(i)}$and hence brings about a reduction of the influence of $V_{ext(i)}$.

Though, according to Sick (2010) "Clear unidimensional variables help us to form conclusions and make decisions free of confounding interpretations." (p. 23) many researchers have questioned the possibility of a unidimensional measure, especially in achievement test. If two distinct components measured together are highly inter-related/correlated, then they are not likely to be factorially distinct and hence can come out as constituting a 'unidimensional' measure which it is not. To McNemar (1946) cited in Revelle and Zinbarg (2009), though measurement implies that one characteristic is being quantified at a time, "there are some domains that consist of related (possibly even highly related) yet discriminable facets that differ even in their true scores. Such domains are themselves not unidimensional, and so it would be unrealistic to expect measures of them to be unidimensional" (p. 152).

A score made on a unidimensional measure is 'clean' and objectively interpretable. In other words, a test score is validly interpretable only to the extent that it is the variance of the construct under measurement that underlies the variability of all its items. Consciously or unconsciously two distinct traits can be measured together, like in the case of language and mathematics in worded mathematics items. If both are prominent, they will act as two distinct factors each with a significant systematic variance. To reduce this effect, worded mathematics items should be expressed with words or language that is simple enough for all the testees so that language does not constitute a disadvantage or an advantage to any testee or group of testees. Since it is mathematics ability that is being measured, the language dimension or level should be kept constant or equalized for every testee. The trait under measurement should be the source of any significant systematic variation in the item variance, while language may constitute a source of insignificant systematic variance. The meaning given to a score on a non-unidimensional measure is questionable. If such score is put to use in any examination-score-based decision, data analysis during research study, policy formulation, etc., such decision might be unfair and biased, the findings of such research cannot reflect the truth which the research study was

searching for, nor would such policy if implemented lead to the solution of the relevant problem. Hence, unidimensionality of a measurement tool is fundamental to any scientific effort at educational measurement.

The second assumption for all IRT models applied to data from dichotomously scored achievement items is local independence. Two measures are said to be independent if their scores do not correlate. Not that two item scores should not correlate, but they should correlate locally, that is, their correlation should be only as a result of what they share with all the items in the test, that is, what the test was designed to measure. If the variance of what they share is removed or held constant, then they should not correlate (see Formula 12). Formally, according to Inaerm (2006), "if the latent variable is maintained at a fixed level, then all the items are independent." (p. 162). Given that what underlies responses to all the items is $\boldsymbol{\theta}$, then:

$$r_{i_1 i_2}/\theta = 0 \qquad (12)$$

In other words, when $\boldsymbol{\theta}$ is held constant, scores from Items 1 and 2 do not correlate. As illustrated in Figure 2, given a three-item test with Items 1, 2 & 3, the common variance they

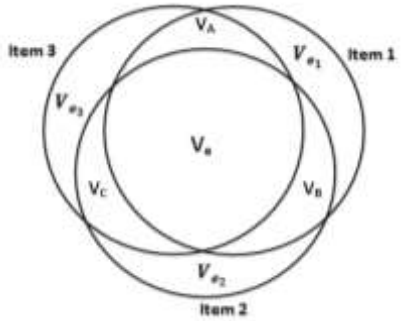

Figure 2: A Venn-diagram illustration of variance component given a three-item test.

Item 1 variance = $V_\theta + V_A + V_B + V_{e_1}$

Item 2 variance = $V_\theta + V_B + V_C + V_{e_2}$

Item 3 variance = $V_\theta + V_A + V_C + V_{e_3}$

share is represented by V_θ [that is, V_{com}, or $V_{trait(i)}$] and this represents the influence of the trait or ability under measurement. While V_{e_1}, V_{e_2} and V_{e_3} represent the random error variance for Item 1, 2 & 3 respectively. V_A, V_B and V_c represent extraneous variance components from systematic extrinsic or biasing factors.

Unidimensionality has to do with the dominance of V_θ given the variance of all the items, V_o; while the second assumption is saying that Items 1 & 2, for example, are locally independent if and only if the correlation between scores on them after V_θ has been controlled or covaried out, is not significantly different from zero. In other words, the correlation between Items 1 & 2 is only that brought about by V_θ, the variance that is common to both of them, the correlation due to extraneous or biasing factors is not significantly different from zero. Technically, IRT insists on items meeting this assumption because, for one, its parameter estimation processes involves the multiplication of probabilities which is only mathematically possible if the events involved are independent. It also makes these confining operational demands to ensure that scores that emanate from testing can be validly interpreted as representing only the value of the trait or behavior under measurement possessed by the testees. Only on such scores are valid decisions or policy formulation possible.

Conclusion

Item response theory (IRT) is an item-based theory that models, in a probabilistic term, the result of the interaction between testee's ability and an item cognitive demand during testing. That is, it predicts the probability of a successful or unsuccessful overcoming a test item as being a function of the difference between the amount of ability possessed by the testee and that demanded by the item before it could be overcome.

Classical test theory (CTT), on the other hand, postulates that the score we observe for a testee on a test, or raw score, is a component of two scores, the true and random error score. In other words, underlying the observed score variance are two independent variances; one a systematic variance and the other a random variance. That associated with the systematic variance is the true score and that which underlies the random variance is the random score. But reality shows that there is more than one source of systematic variance in a testing situation, a truth which CTT often overlooks, and hence generate observed scores loaded with repeatable influences other than that which the test was designed to measure. Hence, CTT-based tests, though they might be highly reliable, are rarely valid, except if its items are developed, administered and scored to allow for the influence of one ability only. IRT, on the other hand, builds this into its test through its assumption of undimensionality, through which it controls for the influence of systematic variation other than that emanating from the ability which the test was designed to measure. Hence, a test that meets IRT assumptions are often found to be valid as a result of controlling for the influence of other systematic sources of variance.

References

Biesheuvel, S. (1974). The nature of intelligence: Some practical implication of its measurement. In J. B. Jeffery, *Culture and cognition: Readings in cross-cultural psychology*. London: Methuen

Harvill, L. M. (1991). Standard error of measurement. *ITEMS. Instructional Topics in Educational Measurement*, Summer, 33-41.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139 – 164.

Inserm, B. F. (2006). The unidimensionality of a psychiatric scale: A statistical point of view. *International Journal of Methods in Psychiatric Research, 8*(3), 162 – 167.

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

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

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Lord, F.M., & Norvick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*(4), 289–374

Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from person's responses and performance as scientific enquiry into score meaning*. Princeton, N. J.: Educational Testing Service.

Nenty, H. J. (1985). Fundamental of educational measurement. Unpublished monograph, University of Calabar, Nigeria.

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. (2015). Conjugal relationship between research and measurement A Keynote Address given 2015 EARNiA Conference in Cameroon.

Nenty, H. J., & Umoinyang I. E. (2000). *Principles of test construction.* Institute of Education, University of Calabar: Helimo Associates.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the Glb: Comments on Sijtsma Psychometrika, *74*(1), 145–154. Retrieved from http://www.personality-project.org/revelle/publications/rz09.pdf

Sick, J. (2010). Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement. *Shiken: JALT Testing & Evaluation SIG Newsletter, 14*(2), 23-29. Retrieved from: http://jalt.org/test/PDF/Sick5.pdf

Stanley, J. C. (1971). Reliability. In R. Thorndike, (Ed.), Educational measurement (2nd ed.) (pp. 356-442). Washington, D. C.: American Council on Education.

Warm, T. A. (1978). A primer of item response theory. National Technical Information Services Department of Commerce, Oklahoma City, OK: U.S. Coast Guard Institute.

Wright. B. D. (1967). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service.