LWATI: A Journal of Contemporary Research, 11(3), 36-44, 2014
ISSN: 1813-2227

From Classical Test Theory (CTT) to Item Response Theory (IRT) in Research Instrument Validation: An introduction to a desirable transition

Idaka Idaka* and Etta Idaka**
**Department of Educational Foundations, Guidance& Counseling, Faculty of Education, University of Calabar, Nigeria*
***Department of Curriculum & Teaching, Faculty of Education, University of Calabar, Nigeria*
Email: idaka.idaka @ yahoo.com
08037238985

ABSTRACT

This paper recognizes the contributions of Classical Test Theory (CTT) in sustaining the validation of psychological instruments. However, consequent upon the well-known pitfalls of the CTT; and sometimes the misleading inferences arising from poorly validated instruments, this paper is a call for a transition from classical test theory (CTT) to item response theory (IRT), or at best an integration. The paper opines that if the negative consequences of misleading research findings are to be avoided in the educational sector, then this transition is inevitable.

Keywords: Instrument validation, Classical test theory, Item response theory.

INTRODUCTION

The collection of data is an extremely important part of all research endeavours, because the conclusions of a study are based on what data revealed. This presupposes the existence of an instrument whether constructed, adopted or adapted. Of course, every instrument, no matter what kind, if it is to be of any value, must allow researchers to draw accurate

conclusions about the characteristics of the people, objects or things being studied.

Psychological instrument e.g. questionnaires are often used to measure abstract qualities that cannot be seen, such as intelligence, honesty, dominance and so on. The relevant questions are: How do researchers evaluate their instruments? How do we know whether such instruments are actually providing accurate information about the characteristics of interest or do we just generate haphazard feedback that sounds plausible? In other words, if an instrument is to be considered useful and accurate, it should meet certain standards that have been set by the psychometric community throughout the years.

What is validation in Research?

Validation is the process through which researchers assess the quality of a psychological instrument by testing such tool against the different standards. Validation therefore asks two basic questions:

(1) How valid is the instrument? In other words, researchers want to know whether the instrument measures accurately. The more that instrument measures what it purports to measure, the more valid the instrument is.
(2) How reliable is the instrument? In other words, researchers want to know whether the instrument measures in a consistent and dependable way. If the results from an instrument contain a lot of random variation, it will be considered less reliable.

Now let us use this simple analogy to paint a picture of the relevance of validity and reliability. Imagine that the very first time you had enough money to fill the tank of your newly acquired automobile; the fuel gauge rather indicated "half tank". After driving for a week, you decided to fill the tank once more and the gauge still indicated 'half tank', your fuel gauge is reliable for being consistent but it is not valid because it is not measuring (gauging) the way it is designed to measure the quantities of fuel in the tank. In the vein, an instrument can be reliable but not valid!

It is however germane, to note that no psychometric tool is perfectly reliable or perfectly valid. All psychological instruments are subject to various sources of error. Hence, reliability and validity are matters of degree on a continuum, rather than reliable/unreliable and valid/invalid on dichotomous scales. It is therefore, more appropriate to ask: "How reliable or valid is your instrument" than is this instrument reliable or valid?

There are two theories that address measurement problem associated with instrument construction. These theories are (i) Classical test theory

(CTT) and (ii) Item response theory (IRT). Both CTT and IRT enable us to predict outcomes of psychological measures by identifying parameters of item difficulty and the ability of testees. They are both concerned with improving the validity and reliability of psychological instrument and provide measures of validity and reliability.

Classical Test Theory

In Nigerian and among education researcher, CTT is the most popular of the two. This theory is regarded as the "true score theory, It introduces three concepts: observed score (test score), true score and error score, which are presented in the form of an equation, linking the observable score (x) to the sum of two unobservable (latent) variable, true score (T) and error score (E). Mathematically, $X = T \pm E$.

The theory assumes that each respondent has a true score which would be obtained if there were no errors in measurement. But unfortunately, measuring instruments such as questionnaires, tests, interest inventory, etc are hardly perfect, hence the observed score may differ from a respondent's true ability. The difference between true score and the observed score is as a result of error in measurement. The error could be random or systematic causing the observed score to be higher or lower. This implies that research instruments are simply fallible and imprecise tools (Joshua, 2005; Magno, 2009). In other words, the observed score is almost always the true score affected by some degree of error.

Validation in CTT

Methods of Validity

As earlier noted validity refers to the extent to which an instrument measures what it is designed to measure. In other words, it is the extent to which an instrument measures accurately. In CTT, there are three methods often used to evaluate the validity of a research instrument. These are; (i) Content validity (ii) construct validity and (iii) criterion related validity

Content Validity

This refers to researchers' subjective assessment of the presentation and relevance of the research instrument in terms of clarity, appropriateness and representativeness of items. Some authors have commented on the status of

content validity in research and are rather in support of jury opinion (experts) in its assessment.

Construct Validity
Construct Validity refers to the extent to which the instrument adequately mirrors the psychological construct that it purports to measure. Construct validity therefore seeks to find out whether those respondents who score high in the scale for instance manifest in real life situation, the construct underlying such instrument. For example, those who scored high in attitude scale of entrepreneurial studies should actually exhibit positive attitude towards entrepreneurships than these who scored low.

Criterion-related Validity
This involves establishing an empirical relationship between scores obtained with a given instrument and some external measure referred to as criterion. Criterion related validity is of two kinds depending on the time frame surrounding the criterion. Consequently, we have concurrent validity and predictive validity.

Methods of Reliability in CTT

According to CTT, reliability is a situation where observed score is consistent from one administration of the instrument to another. In other words, it is the extent to which the instrument measures consistently. There are different ways of establishing reliability of any instrument. These are (i) Test-retest reliability (ii) equivalent from and (iii) internal consistency reliability.

Test-retest reliability: This method is used to assess the consistency of an instrument from one administration to another (i.e. across time). The two sets of scores are then correlated and the coefficient of correlation becomes an estimate of reliability often referred to as the coefficient of stability.

Equivalent form: This method involves the use of two or more equivalent forms of a given instrument. One form is administered to a group and an equivalent form is also administered to the same group. The two sets of scores are then correlated to give coefficient of equivalence which is an index of reliability.

Internal consistency reliability: This method provides an estimate of internal consistency of the items in an instrument. Usually, the instrument is administered once, but at the point of scoring it is split into two halves. The two sets of scores are then correlated to give an estimate of internal consistency, which is an estimate of reliability; which of course, is based on the assumption that items measuring the same construct should correlate.
In the split half method of internal consistency, the coefficient of reliability is then corrected by applying the Spearman-Brown Correction formula; other methods of internal consistency are the Cronbach Alpha, Kuder –Richardson 20 and 21.

Advantages of CTT

CTT has some benefits which has made many traditionalists among educational researchers to continue to patronize it. These are:

(i) In CTT, smaller sample sizes are needed for analysis;
(ii) Simpler mathematical analyses are involved;
(iii) Parameter estimation is straight forward and analyses do not require strict goodness of fit studies to ensure a good fit of model to test the data.

Problems associated with CTT

In spite of the popularity of CTT among Nigerian educational researchers, it is not without some limitations. These limitations are:

(i) In the CTT model, indices such as difficulty, discrimination and stability depend on the characteristics of a sample of individuals to which the test is applied;
(ii) That the scores of the individual test items will be on linear scale for all individuals, whereas it is in the form of a curve;
(iii) The model also assumes that the scores that represent the ability must be in a linear function steadily, hence if the scores of the individual increases in the test, the amount of his ability must be increasing also. But we know that some individuals with high ability sometime get low scores on the test and vice versa.

Consequent upon these limitations associated with CTT which cause inaccuracy in methods and tools of measurement, there was need therefore to develop a method of measuring behaviour in a manner similar to what obtains in the physical sciences. This gave rise to the Item Response Theory (IRT) (Qasem, 2103).

Item Response Theory (IRT)

The IRT goes by different names such as Latent Trait Theory, Strong True Score Theory or Modern Mental Test Theory. Unlike the CTT, IRT focuses on item; it models the response of every respondent to every item in an instrument. It is a statistical theory about the item performance and the abilities that are measured by the items. IRT is a body of theories describing the application of mathematical models on data from questionnaires and tests as a basis for measuring abilities, attitudes perceptions, etc. Basically, the idea here is that the probability of getting an item correct is a function of a latent trait a ability. In other words, a person that possesses the required ability demanded by an item is likely to correctly respond to that item.

Assumptions of IRT

The mathematical models in IRT determine the relationship between an examinee's performance on a test and the ability behind his/her performance, and this mathematical model is the equations which connect the examinee's ability and the probability of getting the correct answer. The three basic assumptions of IRT are:

(i) Unidimensionality - this means that items in a test should measure only one ability or trait;
(ii) Local independence which means examinees responses to any pair of item are statistically independent other things being equal;
(iii) Normal ogive- which says that if item difficulty is plotted against the latent traits of examinees the resultant curve should look like a normal ogive, otherwise referred to as item characteristic curve(ICC)

Methods of Validation in IRT

In IRT, the meaning of validity and reliability differ from CTT since the IRT focuses on the items. Validity therefore, refers to the extent to which individuals or examinees and items have a good ranking in the ability which the test measures. In other words, validity is the ability of any test to rank order the examinees according to their ability and the items according to their level of difficulty (Hambleton, 1983; Qasem, 2013). On the other hand, reliability in IRT refers to the extent to which the measure is independent (free) from groups (Samples) and also from the test items. In other words, the characteristics of the items are not affected by the group that took the test;

and if many versions of the test are given to the same group, they must get the same score and same ranking (Lord, 1968; Qasem, 2013).

There are three models to evaluate the validity and reliability of items in a given instrument based on the three parameters. These are:

(i) The ability of the examinee, (ii) level of difficulty of the items and
(ii) the item ability to discriminate.

Usually, the assumption is that each examinee responding to a test item possesses some amount of underlying ability. Thus, one can consider each examinee to have a numerical value, a score that places him/her somewhere on the ability scale. This ability score is denoted by theta, (Q). This is often represented on an item characteristics curve, which indicates the probability that an examinee with the ability required by an item will give a correct answer to the item. Hence, this probability will be small for examinees with low ability and big for those with high ability.

Advantages of IRT

(i) IRT estimates of item difficulty do not change from one sample to another
(ii) Difficulty indices are also more stable from one form of test to another;
(iii)IRT internal consistencies are stable from one sample to another;
(iv) IRT has significantly less measurement error when compared with CTT.
Other benefits include:
*IRT is very useful when multiple set of items are administered to students in an assessment.
*IRT is used immensely on large-scale testing programmes, especially in achievement and computerized adaptive testing.
*IRT is also very useful in building item banks with the items scaled to different level.
*It useful in the development of criterion- referenced tests and serves as the theoretical foundation for the measurement of personality and psychopathology.

Disadvantages of IRT

The major issue is the complexity of the procedure required in IRT;
Secondly, IRT requires sophisticated statistical techniques for its analysis;
Thirdly, the statistical packages required are not easily available.

RECOMMENDATIONS

(i) It is therefore recommended that educational researcher using CTT should not rely on previous reliability estimates but to estimate their own and indicate any observed differences;
(ii) IRT approaches should be vigorously taught to stakeholders;
(iii) IRT software packages should be made more accessible to intended users;
(iv) More efforts should be directed at teaching students the IRT approaches;
(v) IRT software should be freely distributed by institutions such as NECO, JAMB and other experts.

CONCLUSION

It must be acknowledged that CTT has sustained instrument validation for long and is likely to remain among our researchers in the foreseeable future. However, considering the advantages of IRT, which solves the problems of repeated analysis of data sets every time an instrument is administered in order to re-validate such data, inter alia, this transition is desirable or at least, an integration.

REFERENCES

Anastasi, A. and S. Urbina (2002). Psychology testing. Practice Hall: New York.

Carola, K. and A. Winstertein (2008). Validity and reliability measurement instrument used in research. American Journal of Health System and Pharm, 65.

Hambelton, R. K. (2000). Emergence of item response modeling in instrument and data analysis. Newbury Park CA: Sage

Hembleton, R. K., H. Swaminathan and H. J. Rogers (1991). Fundamentals of items response using IRT. Nwebury Park CA:Sage

Joshua, M. T. (2005). Fundamentals of test and measurement in Education. Calabar: University of Calabar Press.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum Everything Disc.

Magno, C. (2009). Demonstrating the difference between CTTand IRT using derived test data. The International Journal of Education and Psychological Assessment 1 (1): 1-11.

Oluwatayo, J. A. (2002). Validity and reliability uses in Educational Research. Journal of Educational and Social Research 2 (2): 50-57.

Qasem, M. A. N. (2013). A comparative study CTT and IRTan relative to various approaches of evaluating the validity and reliability of research tool. Journal of Research and Method in Education 3(5): 77-81.

Ubi, I. O. (2006) .Item local independence, dimensionality and trend of candidates'

Mathematics performance in University Matriculation Examination in Nigeria. Unpublished PhD Thesis, University of Calabar, Calabar.