TRANSFORMATION FROM CLASSICAL TEST THEORY (CTT) TO ITEM RESPONSE

THEORY (IRT) IN RESEACH

INSTRUMENT VALIDATION

EFFIONG OJI EKPO, IGIRI CHARLES EGBONYI
CROSS RIVER STATE COLLEGE OF EDUCATION
AKAMKPA
+2348067230952

PROFESSOR S. W. BASSEY
CROSS RIVER STATE UNIVERSITY OF TECHNOLOGY
CALABAR

ABSTRACT

The paper deals with the transformation from classical test theory (CTT) to item response theory (IRT) in research instrument validation. The paper recognizes the contribution of classical test theory (CTT) in sustaining the validation of psychological instruments. consequently upon the well – known pitfalls of the CTT and sometimes the misleading inferences arising from poorly validated and unreliable instrument, the study calls for a transformation from classical test theory (CTT) to item response theory (IRT) or at best an integration of both. The various methods of validation of IRT were discussed. It was recommended that IRT approaches should be vigorously taught to stakeholders and IRT software packages should be made more accessible to intended users.

KEY WORDS: instrument, validation, classical test theory, item response theory.

Introduction

Collection of data is an important part for research activities because conclusion of the study can be drawn based on the data collected. An instrument whether constructed, adopted or adapted must allow an investigator to draw conclusion about the characteristics of the subject, events or phenomenon being studied. Psychological instrument are often used to measures abstract quantities that cannot be seen such as anxiety, emotion, attitude, intelligences etc.

The questions that come into our mind are:

1. How do investigators evaluate their instruments?
2. How do we know whether such instruments are actually providing accurate information about the characteristics of interest?
3. Do we just generate data anyhow?

However if an instrument is to be consider useful and accurate, such instrument must meet certain standards that have been set by the psychometricans in educational measurement.

THE MEANING OF VALIDATION

Check or prove the validity or accuracy of an instrument by testing such instrument against the different standards.

Validation ask two different questions

1. How valid is the instrument? An investigators want to know whether the instrument measures accurately. The more the instrument measures what it purpose to measure, the more valid the instrument is.
2. How reliable is the instrument? An investigator wants to know how consistent and dependable an instrument measures.

For example, a reliable man for instance, is a man whose behavior is consistent, dependable and predictable-what he will do tomorrow and next week will be consistent with what he does today and what he has done last week. While an unreliable man is one whose behavior is much more variable.

All psychometric instruments are subject to various sources of error. Hence reliability and validity are matters of degree on a continuum, rather than reliable/unreliable or valid/invalid on dichotomous scales.

Legality

Together with the centralized system of education and the society, there is a centralized national college entrance examination, the element of scores only and the early form as family background only remains the deep gap from the educational purpose of making students a better person with lofty ideas and with the executive values in its spiritual ideas and way of thinking. The national college entrance examination is short of rational legality of education and it can hardly make the improvement of education and only partly embody the social equity in practice so a total reformation is necessary.

There are two theories that address measurement problems associated with instrument construction. These theories are: (1) classical test theory (CTT) and (2) item responsible theory (IRT).

The two theories above enable us to predict outcomes of psychological measures by identifying parameters of item difficulty and ability of examinees. They are concerned with improving the validity and reliability of psychometric instruments and provide measures of validity and reliability.

Classical test theory (CTT)

This theory is known as 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. The relationship between the observed score, true score and error is symbolically expressed as:

$X=T\neq E$, where X is the observed score, while T and E are the true and error scores respectively.

The theory is associated with each test an observed score and a true score. Measuring instruments such as test, questionnaire, rating scale and other are hardly perfect, hence the observed score may differ from a respondent's true ability. The difference between true score and observed score is as a result of error in measurement. The error may be random or systematic, causing the observed score to be higher or lower than the true value. This implies that research instruments are simply fallible and imprecise tools (Joshua, 2005, Magno, 2009)

In other word, the observed score is almost always affected by some degree of error.

Validation in CTT

Validity refers to the degree to which a test actually measures the variable it claims/purports to measure.

There are four methods of validating psychological Instruments in CTT- (1) content validity (2) criterion related validity (3) construct validity (4) face validity.

Content validity

Also known as logical validity refers to the extent to which a measure represents all facets of a given social construct. An element of subjectively exists in relation to determining content validity which requires a degree of agreement about what a particular personality trait represents. A disagreement about a personality trait will prevent the gain of a high content validity

Face validity.

Refers not to what the test actually measure, but to what it superficially appears to measure. It assesses whether the test looks valid to the testees who take it, the administrative personnel who decide on its use and other technically untrained observers.

Criterion-related validity

Criterion related validity is also known as empirical validity and it is the degree of effectiveness with which performance on a test or procedure predicts in a real-life situation there are two types of criterion related validity the only difference between the two is time, namely Concurrent validity and predictive validity.

Construct validity

Refers to the degree to which inferences can legitimately be made from the operationalization in your study to the theoretical constructs on which those operations were based.

Construct validity refers to the extent to which the instrument adequately mirrors the psychological construct that it purports to measure.

Construct validity is related to generalizing, from your study context to other people, places or times. Construct validity involves generalizing from your program or measures to the concept of your program or measures.

Methods of reliability in CCT

Reliability is the repeatability or stability of a measuring instrument. That is the extent to which a test measures whatever it is measuring consistently.

There are four major ways of establishing Reliability of any instrument. These are

(1) Test-retest reliability

(2) equivalent form reliability

(3) split half reliability and

(4) internal consistency.

Test-retest reliability

A particular test is given to a group of testees and the same test is given to the same group after a period of time (about two weeks). The two sets of scores from the two administrations of the same test are correlated. The correlated co-efficient becomes an estimate of the reliability, known as co-efficient of stability. The recommended time lapse between the two times of administration of test is two weeks.

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

Split half reliability:

One test is given to a group once. At the point of scoring the one test is then split into two equivalent halves. Each person in the group is scored on each of the two halves of the test thus resulting in two scores for each individual in the group. The two scores are, correlated using Spearman-Brown prophecy formula

$$r_{tt} = \frac{2r_{xx}}{1+r_{xx}}$$, Where r_{xx} = the correlation coefficient of the scores on the half-tests. r_{tt} = the actual estimate of the test reliabity.

Other measures of internal consistency. Apart from split-half method there are other methods of establishing the internal consistency of a given test. These include:

Kuder-Richardson (K-R) 20 methods

Kuder-Richardson (K-R) 21 methods

Cronbanch co-efficient alpha method

Scorer Reliability method.)

K-R 20 is applicable with items that are scored dichotomously (right =1, wrong=0)

$$KR_{20} = \frac{K}{K-1}\left[1 - \frac{\sum pq}{S_x^2}\right]$$

K-R 21 is applicable for dichotomously scored items that have about the same difficulty.

$$RK_{21} = \frac{K}{K-1}\left[1 - \frac{\bar{X}\,[K-\bar{X}]}{KS_x^2}\right]$$

Cronbanch co-efficient Alpha $= \frac{K}{K-1}\left[1 - \frac{\Sigma S_1^2}{S_x^2}\right]$

Scorer reliability: This is the degree of agreement between different raters' scores of the same test. It is obtained by correlating the scores obtained from two equally competent scorers.

Advantages of CTT

CTT has the following merits

1. 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.
2. Smaller sample sizes needed for analyses
3. Simpler mathematical analyses are involved

Problem associated with CTT

CTT has the following limitation

1. It makes no assumption about matters that are beyond the control of psychometrician.
2. The scores of the individual test items will be on linear scale for all individuals, even when they are really in the form of a curve.
3. The indices such as difficulty, discrimination and stability depend on the characteristics a sample of individual to which the test is applied.

To address the short comings of CTT which cause inaccuracy in methods and tools of measurement, there was need to develop a method of measuring behavior in a manner similar

to what obtains in physical sciences. This gave rise to item-response theory (IRT) (Qasem, 2013).

ITEM RESPOPNSE THEORY (IRT)

In psychometrics, Item response theory also known as latent trait theory, strong true score theory or modern mental test theory's a paradigm for the design, analysis and scoring of tests, questionnaire and similar instruments measuring abilities, attitudes or other variable. Item response theory treat the difficult of each item (the ICCS) as information to be incorporated in scaling items ICC stands for item characteristics curve.

It is the preferred method of developing scales in United States especially when optimal decisions are demanded E.g Graduate Record Examination.

The name item response theory is due to the focus of the theory on the item as opposed to the fest-level focus of classical test theory. Thus IRT models, the response of each Examinee of a given ability to each item in the test. IRT is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of person and item parameters. The main purpose of IRT is to provide a framework for evaluating how well assessments work and how well individual items on assessment work.

Psychometricians use it for developing and designing Examinations maintaining banks of items for Exams and equating the difficulties of items for successive versions of Examinations.

IRT brings greater flexibility and provide more sophisticated information.

IRT provides more sophisticated information which allows a researcher to improve the reliability of an assessment. Under IRT, the primary interest is in whether an Examinee got each individual item correct or not rather than in the raw test scores. When the item response is determined to be correct, the Examinee receives a score of one, an incorrect answer receives score of zero i.e the item is dichotomously scored.

IRT presupposes that for an Examinee to be able to response correctly to items on a test consistency, he/she must possess the ability that is called for by the items. The probability of a correct response to test items therefore depends on the person's trait and the items parameters.

IRT represent a body of related psychometric theories that predicts outcomes of psychological testing such as difficulty of item and the ability of testee.

ASSUMPTION OF IRT

IRT Entails Three Assumptions

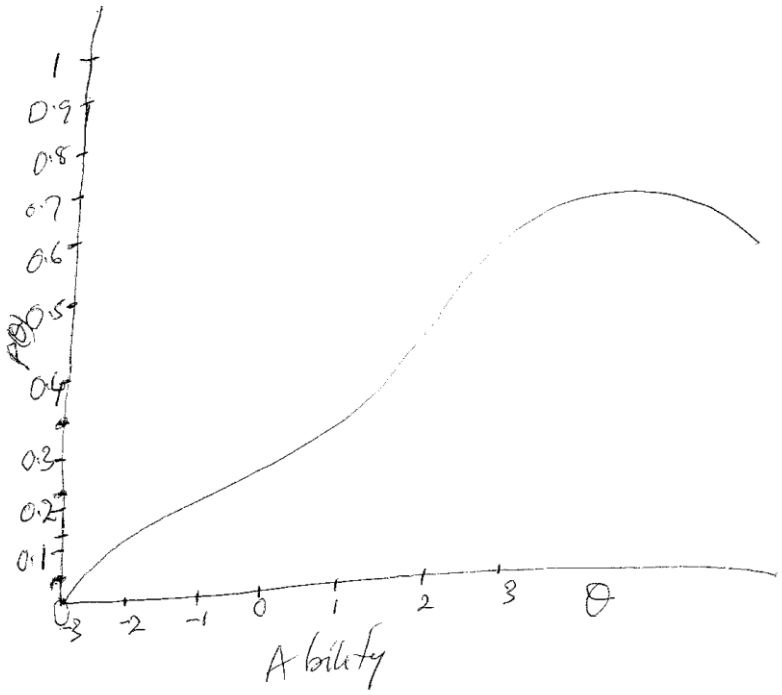
1. A unidimensional trait denoted by θ
2. Local independence of items
3. The response of a person to an item can be modeled by a mathematical item response function (IRF). Unidimensionality should be interpreted as homogeneity, a quality that should be defined or empirically demonstrated in relation to a given purpose. Or use but not a quantity that can be measured.

1. Local independence means (a) that the chance of one items being used is not related to any other items being used and (b) that response to an item is each and ever testee's independent decision, that is there is no cheating or pair or group work.

The IRT gives the probability that a person with a given ability level will answer correctly. Person with lower ability have less a chance, while person with high ability are very likely to answer correctly, that is know-correct assumption.

2. **Unidimensionality:** Here it is believed that any item developed should test one area of knowledge and nothing else, that is items in a test should measure only one ability or trait.

3. **Normal Ogive:** It is believes that when a graph of the testee's ability is plotted the characteristics curve resulting from this will resemble a normal ogive with the steepness showing the difficulty.



The item characteristics curve (ICC) is the building block of item response theory all other constructs of the theory depend upon this curve.

There are two technical properties of an item characteristics curve V_{I3} the difficulty of the item, which describe the functions along the ability scale. It is also known as location index, that is an easy item functions among the low-ability testee's and a hard item functions among the high-ability testee's.

Discrimination index: This describes how well an item can differentiate between testee's having abilities below the item location and those having abilities above the item location. In ICC the steeper curve the better the item can discriminate. The flatter the curve the less the item is able to discriminate.

METHODS OF VALIDATION IN IRT

In IRT, the meaning of validity and reliability differ from CTT, since IRT focuses on the items.

Validity refers to the extent to which testee's an item have good ranking ability which the test measures. That is validity is the ability of any test to rank order the testee's according to their ability and the items according to their level of difficulty (Hambleton 1983:Qasem 2013)

Reliability in IRT is the extent to which measure is independent (free) from groups (samples) and also from the test items. That is the characteristics of the items are not affected by the group that took the test: any if any versions of the test are given to the same group, they must get the same score and ranking (lord, 1968: Qasem 2013)

The three models for the assessment of the validity and reliability of items in a given instrument are based on:

(1) The ability of the testee

(2) Level of difficulty of the items

(3) The item ability to discriminate.

For that ability of the testee, its probability will be denoted by θ. In the case of a typical test item, this probability will be small for testee of low ability and large for testees of high ability. If one plotte θ as a function of ability, the result would be a smooth, s-shape curve which is represented in Item Characteristic Curve (ICC). Which indicate the probability that a testee with the ability required by an item will give a score of answer to the item and receives a score of one, an incorrect answer receives a score of zero i.e that item is dichotomously scored, while this probability will be small for testees with low ability and big for those with high ability.

MERITS OF IRT

IRT have the following merits over CTT

(1) IRT provides significantly greatly flexibility in situation where different samples or tests form used.

(2) IRT findings are foundations for computerized adaptive testing.

(3) IRT is useful in multiple set items for student assessment.

(4) IRT is use for test construction/development such as criterion – reference tests.

(5) IRT parameters are generally not sample or test dependent.

(6) IRT make stronger assessment than CTT

(7) IRT is used in building of item banks.

(8) IRT provides several improvements in scaling items and people, the most model scale the difficulty of item and the ability of people on the same metric.

DEMERIT

(1) The complexity or the procedure in IRT

(2) IRT requires sophisticated statistical techniques for its analysis.

(3) The statistical packages required are not easily available.

One hypothesis was formulated to guide the study, HO:1. There is no significant difference among gender and location on chemistry student's academic achievement.

One instrument was constructed and that is chemistry achievement Test (CAT), the instrument was given to expert in chemistry for validation. A table of specifications was constructed to cover 5 topics for 60[items) multiple choice test this establish content and face validity.

The co-efficient of consistency was determined using mini step software for the chemistry achievement test (CAT). The reliability estimate was 0.89 see table II on the appendix. Nunally (1987) in the standard of reliability coefficient argued that a coefficient of 0.5 would suffice in the early stage, researches in psychological constructs. Using this statistics as a criterion, it could be argued that the estimate met the criterion of stability. Two (2) – way Analysis of variance ANOVA was used to analysis the date collected..

Descriptive statistics

Dependant variable: Achieve = A

Table III

M=1, F=2, URBAN=3, RURAL=4	MEAN	SD	N
1	47.89	11.97	26
2	45.00	9.59	24
3	48.73	9.97	26
4	44.04	11.71	24

Test of between – subjects Effects

Dependent variable: Achieve =A

Table IV

SOURCE	TYPE III SUM OF SQUARE	DF	Mean square	F	SIGN
Gender	17.744		17.744	151	.699
Location	87.423	2	187.423	1.59	.213
Gender and	99.63	2	99.631	847	.362
location		46			
Total	113822.00	50			

Discussion

There is no significant different among gender and location on chemistry students academic achievement. From the descriptive statistic table(III) above it shows that male students has a mean of 47.84 and SD of 11.77 as against their female counterpart with a mean of 45.00 and SD of 9.59,WHILE urban students has a mean of 48.73 and SD of 9.97 and the rural students has a mean of 44.04and SD of 11.71, using 2- way analysis of variance (ANOUA) as shown in table IV.

Gender and location have a f-value of 0.847 as against; sign' of 0.362 at 0.05 level of significance. This implies that the H_o, there is no significance difference among gender and location on chemistry students academic achievement is rejected. That is there significant difference among gender and location on chemistry student's academic achievement.

EDUCATIONAL IMPLICATIONS

1. Find ways of improving students reasoning is a great interest for educators, policy makers.
2. Identified institution strategies that can be use to promote the development of epistemic congrution.
3. Show respect for students assumptions regardless the development stages their exhibits.
4. Create many opportunities for students to analyses other points of view for their evidentiary adequacy to develop and defend their own point of view about controversial issues.
5. Teach students strategies for systematically gathering data, assessing the relevance of the data evaluating, data sources and making interpretive judgments base on the available data.

6. Give students feedback and provide both cognitive and emotional support for their efforts.
7. Encourage to practice their reasoning skills in many settings from other classes to their practicum sites.
8. Help students address issues of uncertainly in judgment making and to examine their assumption about knowledge and how it is gained.

RECOMMENDATION

- ✓ IRT software should be distributed by institutions such as WAEC, NECO, JAMB etc.
- ✓ More other effort should be directed at teaching students the IRT approaches.
- ✓ IRT software packages should be made more accessible to intended users.
- ✓ IRT approaches should be taught to stakeholders.

CONCLUSION

It must be acknowledge that CTT has sustained instrument validation for a long time and is likely to remain among our researchers in the foreseeable future.

However, considering the merit of IRT over CTT, which can solves the problems of repeated analysis of data set every time an instrument is administered in order to re-validate such, interelia, this transition is desirable at least.

REFERENCES

Anastasi, A. & Urbina S. (2002). Psychologic Testing. Practice, Hall: New York.

Carola,K & Winstertein , A. (2008). Validity and reliability measurement instrument used research . American journal of Health system and pharm, 65. (A): sage.

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

BAKER, F.B .(2001). The Basics of item Response theory University of Wisconsin.

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

Lord, F.M (1980). Application of item response theory to practical testing Problems. Hill scale, NJ: Erlbaum everything Disc (2008).

Magno, C,(2009). Demonstrating the difference between CTT Using derived test data. The international journal of education and psychological Assessment 1(1), (I – II).

Oluwatayo , J. A (2002). validity and reliability Users in Educational Research. Journal of Educational and social research 2 (2).5O – 57).

Qasem, M.A .N (2013). A Comparative study CTT and IRT an relative to various approaches of evaluating the validity and reliability of research tools. Journal of research and method in Education 3, (5):77 -81).

UBI.1.0 (2006). Item local independence, dimensionality and trend of candidate's Mathematics performance in University matriculation Examination in Nigeria. Ph.D Thesis, University of Calabar.

MC. VEY, P.J (1972). The reliability of examination in electronic engineering. Report TR24.

Unpublished manuscripts, Department of Electronic and electrical Engineering, University surveys.

Mheren. W.A (1978) Measurement and Evaluation in Educational and psychology. HOLT: Rinehart and Winston.