Item Response Theory as a Basis for Measuring Latent Trait of Interest

By

Adebule Samuel Olufemi
Research Article

Item Response Theory as a Basis for Measuring Latent Trait of Interest

Adebule Samuel Olufemi
Faculty of Education, Ekiti State University, Ado-Ekiti, Nigeria.
Email: doctorolufemiadebule@yahoo.com

ABSTRACT

Item Response Theory (IRT) is one of the most important developments in testing today. It then becomes necessary for all experts in tests, measurement, assessment and evaluation to be well-abreasted of this approach that builds on the concept of item analysis. This write up examines item response theory framework, its merits and demerits; and concludes that Nigerian psychometricians need and require skills and experience in the principle and assumption of IRT.

Keywords: Item, tests, item response test, classical test theory, latent trait, item characteristic curve, Rasch model.

INTRODUCTION

Psychometricians have developed many tests of diverse and imaginative purposes that the early testing pioneers might never have anticipated. Also, from birth to old age a person encounters tests at almost every turning point in life. Psychological tests sum up performance in numbers or classification. Tests measure individual differences in traits or characteristics that exist in some vague sense of the word. According to Thorndike (1918) in Gregory (2006), whatever exists at all exists in some amount. He went further to declare that anything that exists in amount can be measured.

The performance of each testee or examinee is interpreted in reference to a relevant standardization sample or population in a norm-referenced test while a testee’s or an examinee’s stand with respect to very tightly defined educational objectives is determined in a criterion-referenced test (Alonge, 2004).

ITEM RESPONSE THEORY FRAMEWORK

Psychometricians now favour an alternative model of test theory and development known as Item Response Theory (IRT). It is also called the latent trait theory (Baghael, 2008). Lord [1953] stated that an important observation had been made over some years that examinee observed scores and true scores are not synonymous with ability scores. Also, ability scores are more fundamental because they are test independent, where as observed scores and true scores are test dependent.

According to Gregory (2006) and Alonge (1996), when a test is being developed within the item response theory (IRT) framework, the psychometrician posits a single dimension of skill or underlying trait on which all of the test items rely, to some extent, for their correct response. Each testee is hypothesized to have a certain amount of latent trait being measured, what is measured could be verbal proficiency, spatial memory, mathematical reasoning and manipulative skill or even fine-motor skill.

Item Response Test (IRT) builds on the concepts of item analysis.

- The focus is on each item and establishing items that actually measure a particular ability or the respondent’s level of a latent trait.
- IRT rests on the assumption that the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits or abilities.
- In application, the latent trait does not actually exist, because the trait or ability could be a measure on a test, the student’s age, or the student’s grade level.
- Using an item response theory approach, we get an indication of an individual’s performance based not on the Total Score but, rather, on the precise items the person answers correctly.
- IRT involves scaling of items, a procedure that requires quite sophisticated mathematical analyses.
- Item Response Test suggests that the relationship between examinees' item performance and the underlying trait being measured can be described by a monotonically increasing function called an item characteristic function or an item characteristic curve (Hambleton, Swaminathan & Rogers, 1991).

**Construction of Item Characteristic Curve**

1. An instrument developer starts by constructing an item that he believes will measure the ability of interest.
2. The developer gives the item to an appropriate group or groups.
3. Plot the results for each item using an item characteristic curve.
4. Plot a measure of trait or ability (e.g. age/grade level) on the horizontal axis and the probabilities of getting the item correct on the vertical axis.
5. A separate item characteristic curve is graphed for each item based upon a plot of the total test scores on the horizontal axis versus the proportion of examinees passing the item on the vertical axis.
6. He obtains the probabilities of getting the item correct by determining the proportion of persons at different levels who passed an item.
7. The simplest item characteristic curve model is the Rasch model, based upon the IRT of the Danish mathematician, George Rasch (1966).
8. Rasch model makes just two assumptions:
   a. Test items are unidimensional and measure one common trait
   b. Test items vary upon a continuum of difficulty level

![Figure 1: Some hypothetical item characteristic curves](source: Susan (2005))

In the equation or formula below, the position that each test item occupies on the dimension is referred to as the item difficulty (denoted by \( b \)). Also, the position of each testee on this dimension is referred to as his or her proficiency (denoted by \( \theta \)).

An important appealing advantage of IRT model is that the probability of a testee answering a question correctly can be expressed as a precise mathematical equation in terms of both \( b \) and \( \theta \).

\[
P(\theta) = \frac{1}{1 + e^{-(\theta-b)}} \quad \text{or} \quad \frac{e^{(\theta-b)}}{1 + e^{(\theta-b)}}
\]

Where:

i. \( P(\theta) \) is the probability of a testee with proficiency \( \theta \) correctly responding to an item of difficulty \( b \).
ii. \( e \) as a symbol in the equation refers to the base for natural logarithm, which has a constant value of 2.71828.

The formula or mathematical equation was developed by a Danish mathematician called George Rasch in 1960.
Rasch Model and Analysis

Rasch model is one of the psychometric models. It helps us create measures. The most famous application of Rasch model can be found in large education data sets, such as Trends in International Mathematics and Science Study (TIMSS) and National Education Longitudinal Study (NELS) in the forms of students’ academic achievement scores. Rasch model can also create attitudinal scales, such as student engagement level or liking of Mathematics. We all need to be aware that Rasch model can get us good scales only when we design test items and surveys carefully.

Item Response Theory = Latent Trait Theory = Strong True Score Theory = Modern Mental Test Theory

It is a body of theories describing the application of mathematical models of data from questionnaires and tests as a basis for measuring abilities, attitudes, or other variables. It is used for statistical analysis and development of assessments, often for high stake tests such as the graduate record examination. At its most basic level, it is based on the idea that the probability of getting an item correct is a function of a latent trait or ability. For example, a person with higher intelligence would be more likely to correctly respond to a given item on an intelligence test.

Rasch analysis requires a well-defined group of people responding to a set of items for assessment generally. The responses to the items are scored 0 and 1 for two-ordered categories or 0, 1 and 2 (for three-ordered categories) or 0, 1, 2 and 3 (for four-ordered categories) etc to indicate levels of a response on some variables such as health, status, or educational achievement. These responses are then added across items to give each person a total score.

Classical Test Theory (CTT) and Item Response Theory (IRT)

Most educational testing activities in Nigeria today particularly in terms of development, assembling and administration have been based on CTT which evolved with Binet Test a century ago. ITR was developed to overcome the problems with CTT.

CTT is a theory about test scores that introduces three concepts. These are test scores called observed score, true score and error score. A simple linear model X=T+E links the observable test score (X) to the sum of two unobservable (or often called latent) variables.

The main shortcoming of CTT is the circular dependency of items and examinees characteristics. Circular dependency means that test item parameters (characteristics) are defined only with respect to a specific group of examinees, and at the same time, examinees’ characteristics are defined only with respect to a specific test. Here, the item difficulty and item discrimination indices are group dependent, also the observed and true scores are also dependent. They rise and fall with changes in test difficulties. This dependency makes it difficult to compare examinees who took different tests.

According to Hambleton (1989), p and r values are entirely dependent on the examinee samples from which they are obtained. The higher discrimination indices are obtained from heterogeneous examinee sample while lower values are obtained from homogeneous examinee sample. In terms of difficulty indices, higher values will be obtained from examinee sample of above average ability while lower values are gotten from examinee sample of below average ability.

In CTT, there is also an assumption of equal errors of measurement for all examinees (Wikipedia, 2006). IRT provides a foundation in scaling person and item based on response to assessment items. IRT as a body of theories provides a basis for estimating parameters, ascertaining how well data fit a model and investigating the psychometric proportions of assessment.

While comparing, Gregory (2006) in his publication states that some conclusions from classical testing theory do not hold true within the framework of IRT. For example, within the CTT, the standard error of measurement SEM = SD√1−r is assumed to be a constant that applies to all examinees scores regardless of the ability level of a particular testee. But within IRT, the standard error of measurement becomes substantially larger at both extremes of ability. The IRT model concludes that test scores are more reliable for individuals of average ability and increasingly less reliable for those with very high or very low ability (Gregory, 2006).

Another area of contrast pertains to the relationship between test length and reliability. It is almost an axiom in classical test theory that longer tests are more reliable than shorter tests. Of relevance here is the Spearman-Brown prophecy formula used for the adjustment of split-half reliability coefficients. But when IRT models are used, shorter test can be more reliable than longer tests. This is always true especially when there is a good match between the difficulty level of the specific items administered and the proficiency level of the examinee. A good fit between the two parameters allows for a precise or reliable estimate of ability using a relatively smaller number of test items.

However, some benefits obtainable through the use of CTT models include the fact that smaller sample sizes are required for analysis. Also, simpler mathematical analyses are involved, model parameter estimation is
straightforward and analyses do not require strict goodness of fit studies to ensure a good fit of model to test the data. Test developed within an IRT model are better suited to computerized adaptive testing.

**Advantages of IRT**

Item response theory has many advantages when compared with the classical test theory especially when tests are administered by computer. Alonge (1996) explains that in IRT, it is possible to estimate the ability level of a testee with respect to a certain domain of content, using any sets of items of that domain and to estimate the level of item difficulty without reference to the particular testees who take the test.

Also, Beaton & Johnson (1989) states that IRT is frequently used in analyzing test items, in linking test forms and in developing scales to report the results of a national assessment. IRT has a number of advantages when applied to scaling of assessment data. Anderson & Morgan (2008) submits that IRT allows an item to be characterized independently of any sample of items administered to the person. Thus, IRT is very useful when multiple set of items are administered to students in an assessment.

Another advantage of the Rasch model, according to Baghael (2008), is that it builds a hypothetical one-dimensional line along which test items and persons are located according to their difficulty and ability measures. Those items that fall or lie close enough to the hypothetical line contribute to the measurement of the single dimension defined in the construct theory. Those items that fall or lie far from it are measuring another dimension, which is irrelevant to the main Rasch dimension.

Furthermore, it is believed that long distances between the items on the line indicate that there are big differences between item difficulties, so people who fall in ability close to this part of the line are not as precisely measured by means of the test. Other advantages of IRT are:

- Instruments developed using an item response approach are not evaluated the same way as those developed using classical test theory approach.
- IRT approach will theoretically result in an instrument that is not dependent on a norming group. Therefore, the methods for determining standard error of measurement (SEM) are also different. With IRT, the SEM differs across scores.
- Today, IRT is having a significant influence on large-scale testing programmes, especially in the area of achievement and ability adaptive testing.
- IRT is useful in building item banks with the items scaled to different levels. The item banks can provide information that is useful in identifying the academic strengths and weaknesses of the individual. This is possible since the calibrated items can be used to quickly pinpoint the ability level of the person and possibly identify the gaps in knowledge.
- IRT is useful in the development of criterion-reference tests and serves as the theoretical foundation for measures of personality and psychopathology.

**Disadvantages of IRT**

Item response theory has a major disadvantage when compared with the classical test theory. According to Beaton & Johnson (1989), the major demerit associated with IRT is the complexity of the procedure, which requires considerable skill and experience.

- In IRT, the focus is on calibrating each item and examining which items an individual answers correctly, because the focus is on each item, typical measures of score consistency are not applicable.
- Hypothetically, the standard error of measurement (SEM) in an IRT-based assessment is the standard deviation of a distribution of measurement errors that occur when a given population is assessed (AERA, APA & NCME, 1999).
- IRT is complex and involves examining the item characteristics curve of each item, thus, the evaluation of precision or error within the instrument is also quite complex and statistically sophisticated (Susan, 2005).

**CONCLUSION AND RECOMMENDATIONS**

The wind of ICT has blown to all nations of the world and computer is now being used in all schools. Assessment and examination are now being done or administered through the computer. Nigerian experts in test development as well as subject experts need to be aware of the new rules of measurement. There is the need to point out that in addition to straightforward applications of computers, such as presenting test questions, scoring test data and printing test results, computers can be used to design individualized tests based upon real-time feedback during testing, interprets test results according to complex decision rules, write lengthy and detailed narrative reports and present test stimuli in engaging and realistic format.
However, when the prerequisite skills and experience required for a successful use of IRT are not available especially in Nigerian schools, the application of CTT is still acceptable.

REFERENCES


