Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 11, November 2013, pg.7 – 11

RESEARCH ARTICLE

ITEM RESPONSE THEORY

Ayushi Pathak, Kaustubh Patro, Manoj Pathak, Mohit Valecha
Computer Science Department, Dronacharya College of Engineering, Greater Noida, Uttar Pradesh, India

ayushi112233@yahoo.in; kaustubh_patro@yahoo.com; manojpathak1992@gmail.com; mvalecha03@gmail.com

Abstract— Item Response Theory is based on the application of related mathematical models to testing data. Because it is generally regarded as superior to classical test theory, it is the preferred method for developing scales, especially when optimal decisions are demanded, as in so-called high-stakes tests.
The term item is generic: covering all kinds of informative item. They might be multiple choice questions that have incorrect and correct responses, but are also commonly statements on questionnaires that allow respondents to indicate level of agreement (a rating or Likert scale), or patient symptoms scored as present/absent, or diagnostic information in complex systems.
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 person parameter is construed as (usually) a single latent trait or dimension. Examples include general intelligence or the strength of an attitude.

Keywords – Item response theory; IRF; graphical analysis; newton raphson; CTT; future prospects

I. INTRODUCTION

Concept of item response theory was initiated in 1970 by psychometricians to resolve the problems & weakness of classical test theory as these tests were unable to assess the true measurement of ability of students. For so many years and till now, various personality and aptitude tests have been developed for assessments of candidates on various levels .These tests are constructed & the test scores are interpreted using classical test models & procedures which measures the ability of test takers.

However, according to psychologists, classical tests cannot measure the true ability of a test taker and cannot measure the characteristics of test items (aptitude questions) such as their difficulty & discrimination parameter, which are very important for construction of aptitude and psychological tests.

II. ITEM RESPONSE THEORY

Item Response Theory overcomes the problem of classical test theory. It uses various models & assumptions to determine item parameters accurately & precisely & these parameters are used to construct aptitude & psychological tests. Tests constructed with items response models measures the true ability of test takers. It uses various complex mathematical functions (regression, differentiation) to carry out item analysis and ability estimation of students .Various graphs are plotted like **Item Characteristic Curve** & **Test Characteristic Curve** to plot ability of test takers & their probability of answering the question correctly & implementing these mathematical functions and plotting graphs manually, then analyzing the outputs becomes very time consuming process for analysts.

In psychometrics, item response theory (IRT) also known as latent trait theory, strong true score theory, or modern mental test theory, is a paradigm for the design, analysis, and scoring of tests, questionnaires, and

similar instruments measuring abilities, attitudes, or other variables. It is based on the application of related mathematical models to testing data. Because it is generally regarded as superior to classical test theory, it is the preferred method for the development of high-stakes tests such as the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT).

III. GRAPHICAL ANALYSIS : ITEM RESPEONSE THEORY

The name *item response theory* is due to the focus of the theory on the item, as opposed to the test-level focus of classical test theory, by modeling the *response* of an examinee of given ability to each *item* in the test. The term *item* is used because many test questions are not actually questions; they might be multiple choice questions that have incorrect and correct responses, but are also commonly statements on questionnaires that allow respondents to indicate level of agreement (a rating or Likert scale), or patient symptoms scored as present/absent. 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 person parameter is called latent trait or ability; it may, for example, represent a person's intelligence or the strength of an attitude. Item parameters include difficulty (location), discrimination (slope or correlation), and pseudo-guessing (lower asymptote).

The IRF gives the probability that a person with a given ability level will answer correctly. Persons with lower ability have less of a chance, while persons with high ability are very likely to answer correctly; for example, students with higher math ability are more likely to get a math item correct. The exact value of the probability depends, in addition to ability, on a set of *item parameters* for the IRF. For example, in the *three parameter logistic* (3PL) model, the probability of a correct response to an item *i* is:

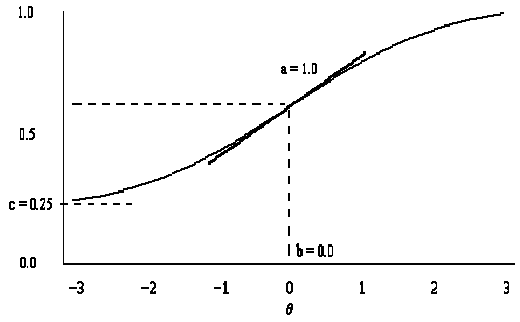
$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}}$$

where θ is the person (ability) parameter and a_i, b_i, and c_i are the item parameters.

The item parameters simply determine the shape of the IRF and in some cases have a direct interpretation. The figure to the right depicts an example of the 3PL model of the ICC with an overlaid conceptual explanation of the parameters. The parameter b_i represents the item location which, in the case of attainment testing, is referred to as the item difficulty. It is the point on θ where the IRF has its maximum slope. The example item is of medium difficulty since b_i=0.0, which is near the center of the distribution. Note that this model scales the item's difficulty and the person's trait onto the same continuum. Thus, it is valid to talk about an item being about as hard as Person A's trait level or of a person's trait level being about the same as Item Y's difficulty, in the sense that successful performance of the task involved with an item reflects a specific level of ability.

The item parameter a_i represents the discrimination of the item: that is, the degree to which the item discriminates between persons in different regions on the latent continuum. This parameter characterizes the slope of the IRF where the slope is at its maximum. The example item has a_i=1.0, which discriminates fairly well; persons with low ability do indeed have a much smaller chance of correctly responding than persons of higher ability.

For items such as multiple choice items, the parameter c_i is used in attempt to account for the effects of guessing on the probability of a correct response. It indicates the probability that very low ability individuals will get this item correct by chance, mathematically represented as a lower asymptote. A four-option multiple choice item might have an IRF like the example item; there is a 1/4 chance of an extremely low ability candidate guessing the correct answer, so c_i would be approximately 0.25. This approach assumes that all options are equally plausible, because if one option made no sense, even the lowest ability person would be able to discard it, so IRT parameter estimation methods take this into account and estimate a c_i based on the observed data.



IRT models can be divided into two families: uni-dimensional and multidimensional. Uni-dimensional models require a single trait (ability) dimension θ. Multidimensional IRT models model response data

hypothesized to arise from multiple traits. However, because of the greatly increased complexity, the majority of IRT research and applications utilize a uni-dimensional model.

IRT models can also be categorized based on the number of scored responses. The typical multiple choice item is *dichotomous*; even though there may be four or five options, it is still scored only as correct/incorrect (right/wrong). Another class of models apply to *polytomous* outcomes, where each response has a different score value.[4][5] A common example of this Likert-type items, e.g., "Rate on a scale of 1 to 5."

An alternative formulation constructs IRFs based on the normal probability distribution; these are sometimes called *normal ogive models*. For example, the formula for a two-parameter normal-ogive IRF is:

$$p_i(\theta) = \Phi\left(\frac{\theta - b_i}{\sigma_i}\right)$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution.

The normal-ogive model derives from the assumption of normally distributed measurement error and is theoretically appealing on that basis. Here b_i is, again, the difficulty parameter. The discrimination parameter is σ_i, the standard deviation of the measurement error for item *i*, and comparable to $1/a_i$.

One can estimate a normal-ogive latent trait model by factor-analyzing a matrix of tetrachoric correlations between items.[7] This means it is technically possible to estimate a simple IRT model using general-purpose statistical software.

With rescaling of the ability parameter, it is possible to make the 2PL logistic model closely approximate the cumulative normal ogive. Typically, the 2PL logistic and normal-ogive IRFs differ in probability by no more than 0.01 across the range of the function. The difference is greatest in the distribution tails, however, which tend to have more influence on results.

The latent trait/IRT model was originally developed using normal ogives, but this was considered too computationally demanding for the computers at the time (1960s). The logistic model was proposed as a simpler alternative, and has enjoyed wide use since. More recently, however, it was demonstrated that, using standard polynomial approximations to the normal *cdf*,[8] the normal-ogive model is no more computationally demanding than logistic models.

One of the major contributions of item response theory is the extension of the concept of reliability. Traditionally, reliability refers to the precision of measurement (i.e., the degree to which measurement is free of error). And traditionally, it is measured using a single index defined in various ways, such as the ratio of true and observed score variance. This index is helpful in characterizing a test's average reliability, for example in order to compare two tests. But IRT makes it clear that precision is not uniform across the entire range of test scores. Scores at the edges of the test's range, for example, generally have more error associated with them than scores closer to the middle of the range.

Item response theory advances the concept of item and test information to replace reliability. Information is also a *function* of the model parameters. For example, according to Fisher information theory, the item information supplied in the case of the 1PL for dichotomous response data is simply the probability of a correct response multiplied by the probability of an incorrect response, or,

$$I(\theta) = p_i(\theta)q_i(\theta).$$

The standard error of estimation (SE) is the reciprocal of the test information of at a given trait level, is the

$$\mathrm{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}.$$

Thus more information implies less error of measurement.

For other models, such as the two and three parameters models, the discrimination parameter plays an important role in the function. The item information function for the two parameter model is

$$I(\theta) = a_i^2 p_i(\theta)q_i(\theta).$$

The item information function for the three parameter model is

$$I(\theta) = a_i^2 \frac{(p_i(\theta) - c_i)^2}{(1 - c_i)^2} \frac{q_i(\theta)}{p_i(\theta)}$$

[18]

In general, item information functions tend to look bell-shaped. Highly discriminating items have tall, narrow information functions; they contribute greatly but over a narrow range. Less discriminating items provide less information but over a wider range.

Plots of item information can be used to see how much information an item contributes and to what portion of the scale score range. Because of local independence, item information functions are additive. Thus, the test

information function is simply the sum of the information functions of the items on the exam. Using this property with a large item bank, test information functions can be shaped to control measurement error very precisely.

Characterizing the accuracy of test scores is perhaps the central issue in psychometric theory and is a chief difference between IRT and CTT. IRT findings reveal that the CTT concept of reliability is a simplification. In the place of reliability, IRT offers the test information function which shows the degree of precision at different values of theta, θ.

These results allow psychometricians to (potentially) carefully shape the level of reliability for different ranges of ability by including carefully chosen items. For example, in a certification situation in which a test can only be passed or failed, where there is only a single "cutscore," and where the actually passing score is unimportant, a very efficient test can be developed by selecting only items that have high information near the cutscore. These items generally correspond to items whose difficulty is about the same as that of the cutscore.

The person parameter θ represents the magnitude of *latent trait* of the individual, which is the human capacity or attribute measured by the test.[19] It might be a cognitive ability, physical ability, skill, knowledge, attitude, personality characteristic, etc.

The estimate of the person parameter - the "score" on a test with IRT - is computed and interpreted in a very different manner as compared to traditional scores like number or percent correct. The individual's total number-correct score is not the actual score, but is rather based on the IRFs, leading to a weighted score when the model contains item discrimination parameters. It is actually obtained by multiplying the item response function for each item to obtain a *likelihood function*, the highest point of which is the *maximum likelihood estimate* of θ. This highest point is typically estimated with IRT software using the Newton-Raphson method.[20] While scoring is much more sophisticated with IRT, for most tests, the (linear) correlation between the theta estimate and a traditional score is very high; often it is .95 or more. A graph of IRT scores against traditional scores shows an ogive shape implying that the IRT estimates separate individuals at the borders of the range more than in the middle.

An important difference between CTT and IRT is the treatment of measurement error, indexed by the standard error of measurement. All tests, questionnaires, and inventories are imprecise tools; we can never know a person's *true score*, but rather only have an estimate, the *observed score.* There is some amount of random error which may push the observed score higher or lower than the true score. CTT assumes that the amount of error is the same for each examinee, but IRT allows it to vary.[21]

Also, nothing about IRT refutes human development or improvement or assumes that a trait level is fixed. A person may learn skills, knowledge or even so called "test-taking skills" which may translate to a higher true-score. In fact, a portion of IRT research focuses on the measurement of change in trait level.[22]

The aim of this research work was to help in developing a fully functional item analysis software system that ensures efficient test data analysis and reduce the time consuming task of carrying out manual analysis. This application validates and cleanses data and stores them back in order to make them readily available for analysis. Analysts can take the output report in the form of hard copy or store them on system for studying them later.

IV. FUTURE PROSPECTS FOR MODIFICATIONS

This software can be modified to add more functionality into it by adding modules to it.

- Currently it carries out analysis based on 2PL model i.e. 2 Parameter Logistic Model, it can calculate only 2 parameters of an item (aptitude question) i.e. difficulty & discrimination factor. It can be modified to carry out analysis based on 3PL model i.e. it will calculate 3 parameters of an item (aptitude question) i.e. difficulty level of an item, discrimination parameter & guessing factor.
- Currently it carries out analysis for dichotomous data (only one correct answer is selected by test taker) only. It can be modified to analyze multichotomous data (where more than one correct answer is selected by test taker).
- Functionality of this software serves as the basis of computer adaptive testing. It can be modified by adding functional modules and interfaces to make it complete computer adaptive testing software.
- Currently only simple excel files are being used to store test data. Data can be stored on remote server and application can be modified suitably to access data from remote server.

REFERENCES

[1] Swaminathan, "ITEM RESPONSE THEORY, principles & applications", nijhoff publishing, 2005
[2] Frank.B Baker , "The Basics of item response theory" , 2nd edition, ERIC clearinghouse publications, 2001
[3] Teodor Danceu and Lucian chirita, "The Definative Guide To JasperReports", JasperSoft Corporation publishing, 2007
[4] K.K Aggarwal, "Software Engineering", 4th edition, pearson publication,2008
[5] IRT statistics
[6] Similar project based on Item response theory in DIPR
[7] en.wikipedia.org/wiki/item response theory

