



COMPARISON OF ITEM RESPONSE THEORY WITH CLASSICAL TEST THEORY OF ASSESSMENT.

1. MBBS, FCPS (Anaesthesia), FCPS (Cardiothoracic Anaesthesia) Assistant Professor Anaesthesia and Intensive Care
Ch. Pervaiz Ellahi Institute of Cardiology, Multan.
2. MBBS, FCPS
Senior Registrar Gynaecology and Obstetrics
Nishtar Hospital Multan.
3. MBBS, FCPS
Consultant Pulmonologist Intensive Care
Ch. Pervaiz Ellahi Institute of Cardiology
4. MBBS, FCPS
Professor Anaesthesia and Intensive Care
Ch. Pervaiz Ellahi Institute of Cardiology, Multan.

Correspondence Address:

Dr. Aamir Furqan
Department of Anaesthesia and Intensive Care
Ch. Pervaiz Ellahi Institute of Cardiology, Multan.
draamir2009@hotmail.com

Article received on:

27/09/2019

Accepted for publication:

25/11/2019

Aamir Furqan¹, Rahat Akhtar², Masood Alam³, Rana Altaf Ahmed⁴

ABSTRACT... Objectives: This article is designed for comparison and contrast of item response theory measurement with classical measurement theory (Classical Measurement Theory) as well as to determine the various advantages offered by item response theory in the setting of medical education. **Summary:** Classical measurement theory is being impartial and inherent, is used more often than other models in medical education. However, there is one restriction encountered in the use of classical measurement theory that is it sample dependent and the data is bewildered in the specified sample that the researcher has assessed. Whereas, the score in item response theory separate from the sample or stimuli of assessment. Item Response Theory is consistent, it allows for easy evaluation of examination scores enabling the score to be placed in constant measurement scale and compare the change in students' ability with time. There are various models of Item Response Theory out of which three are discussed along with their statistical assumptions. **Conclusions:** Item Response Theory being a capable tool is able to simplify a major issue of Classical Measurement Theory, i.e. bewilderment of skill of examinee with item characteristics. The Item Response Theory measurement inscribes the problems in medical education like removing rater mistakes from evaluation.

Key words: Computer-based Testing, Classical Measurement Theory, Examinee, Item Response Theory, Medical Education.

Article Citation: Furqan A, Akhtar R, Alam M, Ahmed RA. Comparison of item response theory with classical test theory of assessment. Professional Med J 2020; 27(3):448-454. DOI: 10.29309/TPMJ/2020.27.3.3453

Classical measurement theory: basic concepts and limitations

A simple method of understanding this theory is through the following example. An anatomy professor is to give a written test to medical students. The questions on the test are all new for the present year but the design of test is to measure the same content of anatomy as previous year's exam.¹ It has come into observation of the professor the present year score of students is significantly less than the previous year students; the present year test of anatomy seems to be harder than last year's test. It leads the conclusion by the professor that present year students of anatomy are less capable and proficient in anatomy as compared to last year's students. Is the conclusion drawn by the professors accurate?

These types of scenarios repeatedly come across

in the setting of medical education. Classical Measurement Theory is usually used for the assessment which leads to the confounding of students' skills with the intrinsic characteristics of the test components. This is true for written, objective or skill-based proficiency assessment.

Most of the educators are well known with the Classical Measurement Theory, although some of them do not realize it. All the data and numerical procedures are derived from Classical Measurement Theory that used for control of quality and evaluation of medical education.²

Before the elaborative discussion on modern theories of measurement, fundamental ideas of classical measurement must be reviewed. There are two parts of every test score which is error and true score. It can be given by a simple formula

which is

$$X = T + e$$

where X is the observed score on the assessment;

T is the true score, and e is random errors of measurement.

These terms need some elaboration which is as follows: The observed score (X) signifies the observed quantity calculated from evaluation. The raw score could be one of the given scores such as ratings given by teaching staff, number of correct answers to objective test, the percent of correct score on an exam. On the basis of Classical Measurement Theory, the observed score is comprised of two parts i.e. the error and the true score. The true score, being the most important part of Classical Measurement Theory measurement needs some discussion. It is usually defined as long-run mean score or average score. It is the type of score that can't be accurately determined or directly observed, it can however be estimated only. It is the mean of all the scores that are taken on a test or precisely equivalent test taken by same student in infinite number of times, repeatedly.

Principally, Classical Measurement Theory is associated with the determination of errors of calculation or true score.³ The error that comes across in Classical Measurement Theory is invariably non-systematic, random error and not associated with the true score. The random error comprised of unchecked and random conditions that interrupt the accurate evaluation of examinee's true skill or competency. The unsatisfactorily formed test components or stimuli, inappropriate testing circumstances or intrinsic states of examinee such as ailment or lack of attention can be regarded as the cause of such random errors. All of them lead to measurement error to the observed score of examinees. Many advantages of Classical Measurement Theory are seen. For instance, it is easy to understand and proceed, it is simple and straightforward. Students either get 1 point on being correct or 0 marks on being incorrect. For most local assessments in medical studies, Classical Measurement Theory

provides the user and student well and is precise in minimizing the decision errors.

The equating methods of Item Response Theory are easier to implement in comparison to Classical Measurement Theory.⁴ All the significant data such as the difficulty of item, item discrimination, that are usually used for evaluation of assessment rely on a specific sample of students being studied upon. This problem of confounding can be overcome by taking into account a larger number of students taking the test as well as ensuring a consistent inherent ability of the examinee over time. If Classical Measurement Theory scores are equated statistically, enable authentic comparability of students' skills that change from year to year.

Item response theory measurement models

A resolution of this confounding issue is the utilization of another model of measurement i.e. Item Response Theory.⁵ With the help of this model, the issue of confounding of students' skills and item difficulty, as well as the dependency on the sample, is resolved as they are theoretically constant. It implies that after statistically fitting a mathematical model to observed evaluation data and meeting all the presumptions, it is possible to estimate the skills of a student independent of particular questions given on test, test items and characteristics.

Item Response Theory belongs to a category of psychometric measurement models in order to determine the ability of examinee on the aspect being evaluated and item difficulty on the same scale. This scale is consistent such that the ability of examinee is determined independently of a specific set of elements given. Similarly, item difficulty is also evaluated independently of the specified sample of examinee taking the tests.

With the advance of computer-based testing, the discipline of Item Response Theory has moved forward in a significant manner.⁶ Now software of Item Response Theory is available for personal computers enhancing the benefits of Item Response Theory measurement. The question here arises if Item Response Theory is such a

wondrous tool, why hasn't it been adopted by evaluation of current testing?

Dimensionality

There is a group of presumptions to be met with Item Response Theory measurement just like all other statistics.⁷ The basic assumption for all models of Item Response Theory used customarily is "unidimensionality". The assessment must rely on distinct, undivided underlying characteristics or construct. For most commonly used models of Item Response Theory, the test is based on more than one construct, which disables the utilization of Item Response Theory model and hence it is incapable of estimating examinee's skills. Various methods are available to determine the dimensionality of the test, the most frequently used is the factor-analytic method by the use of item-level associations.

Local Independence

Characteristics of test items and parameters of a population like discrimination and difficulty are determined through a process named "item calibration". In order to successfully arrange the test questions with the help of Item Response Theory measurement, it is necessary for the items to be independent locally.⁸ It is also essential for Classical Measurement Theory but Item Response Theory model is somewhat resilient to this presumption as compared to Classical Measurement Theory.

Sample Size

For the proper working of Item Response Theory, a larger size of the sample of the examinee is needed. The minimum of which must be comprised of 200 participants. Increase in complexity of model requires a greater size of the sample. The requirement of a larger size accounts for its limited use in the setting of medical education.⁹ But the sample size also appears to be troublesome in Classical Measurement Theory as well for the analysis of items. In order to get stable and reliable indices, the sample size must comprise of 200 subjects.

How Item Response Theory Works

The fundamental presumption of Item Response

Theory is based upon a probability. It is the underlying skill of an examinee with respect to the characteristics being tested as well as the statistical trait of the item being a test that influences the probability of answering a question correctly asked by the examiner on the test.¹⁰ This relationship of ability and answering the question can be demonstrated with the help of a mathematical function termed as "item characteristic Curve" (ICC). If it is desired to pursue the Item Response Theory scaling analysis, an adequate mathematical model should be selected, one that could demonstrate the fitting of data and meet the necessary presumptions.

Common Item Response Theory Models

Three models are currently being used for tests which score as "right or wrong". They are named for the number of traits that they utilize to determine the examinee skills. One model of Item Response Theory is known as "Rasch model". It uses one parameter i.e. item difficulty to determine to student and item characteristics. The skill of examinee can be determined accurately if the evaluation data fit the Rasch model. This model is used widely throughout the world in various settings of medical education.¹¹ As this model needs a lesser number of examinees, this model is most beneficial for medical educators having a greater size of the class.

For greater scales of assessments, two other models of Item Response Theory are also used i.e. the two-parameter and three-parameter models. Another item is added in the two-parameter model i.e. item discrimination. Whereas, further an item is added in the three-parameter model which is "guessing" parameter. These models are being utilized either experimentally or in association with Classical Measurement Theory statistics for various large-scale assessments in North America.

In the language of Item Response Theory model, the ability of examinee is determined on a "theta scale", which usually ranges from -4σ to $+4\sigma$, where 0 is the mean. This scale is seldom used for reporting scores.

The idea of “test reliability” applies on all the assessments, including the ones that are Item Response Theory-based.¹² This concept is enhanced and explored in Item Response Theory measurement by statistics known as “item information functions”. It is a graphical display of the role of test items to the evaluation of examinee skill at multiple levels of skill. Generally, the greater the item discrimination, the greater is the information contribution by that trait to the calculation of skill.

Item Response Theory and computer-based testing

Computer-based testing is getting fame in the modern world. Various high-stakes and large-scale assessments among Europe and North America are presently conducted in the CBT pattern, either in the form of adaptive or linear examination. The linear form utilizes a computer to give questions and then scores the test soon as the examinee finishes the test. It usually gives a fixed form of test i.e. a specified number of questions that are already selected by human test evaluators. Various forms of test can be developed for each test presentation and a particular test form is then selected randomly to be given to an individual examinee.

On the other hand, the computer adaptive test is a specified form of computer-based testing in which the computer selects every test item to be given to the examinee.¹³ Hence, the test “adapts” to the examinee skills as the test advances. The great benefit of adaptive computer-based testing is that by the use of Item Response Theory model, it allows the immensely accurate estimate of examinee skills with the use of shorter tests. Nonetheless, the length of the test is also concluded by the requirements of the test as the adaptive test must meet particular item details to appropriately sample the criteria of knowledge being evaluated. Adaptive testing necessitates a great number of pre-formed test items as the exposure to every item should be composed thoroughly to assure the continued safety of test items as well as the Item Response Theory of test scores. The principle of completely adaptive testing is gradual complexity, i.e. the computer

is programmed to gradually select tougher questions if the examinee is able to answer the simpler questions. It is done until examinee is unable to get the correct answer on a harder level then the computer is designed to provide with a slightly easier question until the Item Response Theory-estimated perplexity becomes equivalent to that to examinee skills.

Various types of computer adaptive testing system are in use. The most complex one being the completely adaptive system as it uses a number of variables simultaneously and it needs to be managed by a computer system in real time. However, if the length is fixed and has pre-formed item coverage such that the examinee is directed by a succession of smaller tests of a variable degree of difficulty.

A significant benefit of computer-adaptive testing is that the skill of examinee can be estimated precisely with the help of smaller test inside the constraints of adequate testing contents. Item Response Theory measurement model is necessary for computer-adaptive testing as each examinee is to get a different set of assessment questions. Hence, the items of assessment must comprise of those item difficulties that are pre-formed by the help of Item Response Theory model so that the computer is able to select those test questions which match the substantial skill of examinee. The quality of consistency of Item Response Theory measurement is necessary for computer adaptive testing.

However, it is still required by the test developer to carefully organize and form the test questions as well as balance the test content having precise particulars needed for the test. In the circumstances where the content of the test is intricate and the test comprises of various disciplines, the need of appropriate sampling of contents might raise the number of items required for appropriate testing even in adaptive testing.

Other Applications of Item Response Theory in Medical Education

Usually, in the setting of medical education, rating scale data is used for various assessments. For

instance, many clinical skill evaluations comprise of rating done by clinical staff, preceptors, teachers or other faculty members. Sometimes, multiple raters are present to give the clinical rating to the examinee. Most of the calculation errors are due to the raters rather than the rating scale or the items rated.

In order to calculate the error of measurement by the raters, various methods are available such as generalizability theory. Item Response Theory model provides not only the means to determine the measurement error but it also enables the compensation for the given error in order to eliminate it.¹⁴

A version of one -parameter model readjusts the data of rating scale with that of rating score divergence that is caused by rater measurement error and fixes the final rating to decrease or remove the rater error. Hence, a primary source of inaccurate rating data is eliminated and legitimacy proof for clinical skill rating is ameliorated. There is software available that brings about these adjustments with the use of expansion of the Rasch model.

Software

The measurement by Item Response Theory model is not only experimental or the fundamental tool used for methodological research. A number of computer software are available commercially that calibrate the tests with the help of Item Response Theory models, which are also cost-effective. As this software is not much user-friendly, there are training workshops that are organized by the software makers and professional corporation.

Issues in Item Response Theory Measurement

Item Response Theory measurement model proves to be a useful tool. Nonetheless, it must be used correctly. Any usage otherwise may bring about more damage than benefit. In order to successfully imply this model, the presumptions of local independence and unidimensionality must be fulfilled to evaluate real data. Both these presumptions are inferable tentatively with the use of association techniques, but in order to

analyze successfully, adequately large samples of examinee must be present.

Sample size also plays an important role in Item Response Theory model.¹⁵ A minimum of 200 examinees must be present to apply the Rasch model to determine the fitting of the model to actual test data. Whereas, for the three-parameter model, about 1000 examinees are required. Although a lesser number of examinees may be used for evaluation of student skill levels, there would be a greater standard error than that of larger samples.

The statistics used to determine the excellence of suitability of Item Response Theory model to data are contentious and precarious particularly for those of two-parameter and three-parameter models. There has been a failure of development of unanimity to which statistics are to be implied or the method of evaluation of the ones that are already in use, or what reforms should be taken in case of misfit data.

The currently active researched field in Item Response Theory model is the new and complex models of Item Response Theory. Most of the presently implied research is Item Response Theory is regarding the administration of Item Response Theory models to computer-adaptive testing.

The leading subjects include types of adaptive delivery, adaptive test safety problems, control of item exposure procedures, item development methods, developing and pretesting bigger quantities of test items needed for adaptive testing, particularly medium difficulty test items.

CONCLUSION

Item Response Theory is a strong evaluation tool which calculates the examinee capability and test and item difficulty on the same scale. Confounding of student skill and item difficulty can be removed provided that the statistical presumptions are fulfilled and the test data fits the Item Response Theory model. Item Response Theory model gives a latest and strong tool to the medical educators to precisely determine

the student skill in the practical field of study. It also enables to eliminate and adjust a significant source of measurement error i.e. rater error with regards to clinical assessment.




The implementation of Item Response Theory model is within the reach of medical educators in the setting of adequate and proper usage. Since the Item Response Theory software is available generally, with the help of some collective endorsement by the experts of Item Response Theory model to medical educators, these evaluation techniques can be favorably implicated in various settings of medical education.

Copyright© 25 Nov, 2019.

REFERENCES

1. Balsis S, Ruchensky JR, Busch AJ. **Item response theory applications in personality disorder research.** *Personal Disord.* 2017 Oct; 8(4):298-308.
2. Birnbaum A. **Some latent trait models and their use in inferring an examinee's ability.** In: FM Lord, MR Novick, eds. *Statistical Theories of Mental Test Scores.* Reading, Massachusetts: Addison-Wesley; 1968: 395-479.
3. Thomas ML. **The value of item response theory in clinical assessment: A review.** *Assessment.* 2011 Sep; 18(3):291-307.
4. Van der Linden WJ, Hambleton RK. **Item response theory. Brief history, common models, and extensions.** In: WJ van der Linden, RK Hambleton, eds. *Handbook of Modern Item Response Theory.* New York: Springer-Verlag; 1997: 1-28.
5. Downing SM. **Item response theory: Applications of modern test theory in medical education.** *Med Educ.* 2003 Aug; 37(8):739-45.
6. Kreiter CD, Ferguson KJ. **Examining the generalizability of ratings across clerkships using a clinical evaluation form.** *Eval Health Professions* 2001; 24:36-46.
7. Kreiter CD, Ferguson K, Gruppen LD. **Evaluating the usefulness of computerized adaptive testing for medical in-course assessment.** *Acad Med.* 1999 Oct; 74(10):1125-8.
8. McLeod L, Swygert KA, Thissen D. **Factor analysis for items scored in two categories.** In: D Thissen, H Wainer, eds. *Test Scoring.* Mahwah, New Jersey: Erlbaum; 2001; 189-216.
9. Beck CT, Gable RK. **Item response theory in affective instrument development: An illustration.** *J Nurs Meas.* 2001 Spring-Summer; 9(1):5-22.
10. Folk VG, Smith RL. **Models for delivery of CBTs.** In: CN Mills, MTPotenza, JJ Fremer, WCWard, eds. *Computer-Based Testing: Building the Foundation for Future Assessments.* Mahwah, New Jersey: Lawrence Erlbaum Associates; 2002; 41-66.
11. De Champlain AF. **A primer on classical test theory and item response theory for assessments in medical education.** *Med Educ.* 2010 Jan; 44(1):109-17.
12. McLeod L, Swygert KA, Thissen D. **Factor analysis for items scored in two categories.** In: D Thissen, H Wainer, eds. *Test Scoring.* Mahwah, New Jersey: Erlbaum; 2001; 189-216.
13. Sébille V, Hardouin JB, Le Néel T, Kubis G, Boyer F, Guillemin F, Falissard B. **Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients--a simulation study.** *BMC Med Res Methodol.* 2010 Mar 25; 10:24.
14. Hambleton RK, Swaminathan H, Rogers HJ. **Fundamentals of item response theory.** Newbury Park, California: Sage Publications; 1991.
15. Chang CH, Reeve BB. **Item response theory and its applications to patient-reported outcomes measurement.** *Eval Health Prof.* 2005 Sep; 28(3):264-82.

AUTHORSHIP AND CONTRIBUTION DECLARATION

Sr. #	Author(s) Full Name	Contribution to the paper	Author(s) Signature
1	Aamir Furqan	Conceive idea, Design study.	
2	Rahat Akhtar	Literature review, Manuscript writing.	
3	Masood Alam	Data analysis.	
4	Rana Altaf Ahmed	Proof reading.	