Establishing Psychometric Properties of the MSU-TCTO Senior High School Entrance Examination Using Classical Test Theory and Item Response Theory

Jeffrey Imer C. Salim, Mindanao State University, Philippines
Wilham M. Hailaya, Mindanao State University, Philippines

The European Conference on Education 2022
Official Conference Proceedings

Abstract
Achievement Testing is widely used in assessing the psychological capabilities of a person. Thus, correct test constructs are important in achieving the purpose of testing. The Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam (SHSEE) is the first MSU-TCTO school-made paper-and-pen achievement test that was conducted on November 18, 2018, to 1,260 students in different schools in Tawi-Tawi and is given annually to prospect senior high school students. It is composed of 75 English, 40 Mathematics, 30 Science, and 25 Aptitude multiple-choice questions. This study aimed to establish the psychometric properties and the level of adequacy of the examination using the Classical Test Theory (CTT) and Item Response Theory (IRT) models, and any significant difference thereof. The study employed a descriptive quantitative design and used the raw data from the research instrument, which is the scored answer sheet of the 200 examinees. Stratified sampling was applied. Statistical Program for Social Sciences (SPSS) was used to determine the reliability indices according to CTT and IRT. The study concluded that the test items of SHSEE were highly adequate and reliable on both CTT and IRT. Furthermore, there is a significant difference between the reliability index under IRT and CTT models at 0.05 level of significance, but not at 0.01, which gave slight inconsistency in the result. The study recommends to the test committee to further enhance the examination and use Item Response Theory as its statistical treatment.

Keywords: Achievement Testing, Psychometrics, Item Response Theory, Classical Test Theory, Difficulty Index

Introduction

In education, certain measurement tools such as achievement tests are used in order to assess if the students have mastered the course content. And based on these test scores, a student's journey will be affected. Thereby, correct test constructs are important for any examination to serve its purpose such as testing the psychological capabilities of a person. The problem of improving and quantifying the psychological measurement is addressed by doing a psychological testing.

The Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam (MSU-TCTO SHSEE) is the first MSU-TCTO school-made paper-and-pen achievement test that was conducted on November 18, 2018, to 1260 students in different schools in the municipalities of the Province of Tawi-Tawi in the Philippines. The test was designed specifically to assess the junior high school students in Tawi-Tawi who aim to enroll in the MSU-TCTO Senior High School.

The questionnaire is composed of seventy-five (75) multiple-choice questions (MCQ) for English, forty (40) MCQs for Mathematics, thirty (30) MCQs for Science, and twenty-five (25) MCQs for the aptitude. The examination is set to assess the mental and psychological capabilities of all students before they are given admission to the senior high school program of the university. And, it is expected to be conducted every year.

This study aimed to establish the psychometric properties of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam (MSU-TCTO SHSEE) for a deeper analysis and possibly improvement of the Standardized Entrance Exam.

It specifically tried to answer the following questions:

1.) What is the level of reliability or adequacy of the test item of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam using the Item Response Theory (IRT)?

2.) What is the level of reliability or adequacy of the test item of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam using Class Test Theory (CTT)?

3.) Is there a significant difference of the reliability or adequacy of each item of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Exam using Item Response Theory and Classical Test Theory?

The results of this study will help in the further improvement of the standardized examinations that will be given by the university to its prospect students, which will guarantee better evaluation and assessment of the test-takers.

1.1 Psychological Testing

Psychological testing has to do with procedures for selecting, administering and interpreting test scores in an applied setting (Maloney & Ward, 1976). Test fairness is indeed a very

crucial social issue. Thus, the psychometric properties of tests which encompasses information regarding the test score biases must always be an aspect that notifies the use of tests in actual situations.

There are various types of psychological testing like intelligence tests (i.e. Stanford-Binet Intelligence Test and Wechsler Intelligence Scales), academic achievement tests (i.e. Scholastic Achievement Tests or SAT and Graduate Record Examination or GRE), structured personality tests (i.e. California Psychological Inventory or CPI and NEO Personality Inventory), and career interest/guidance instruments (i.e. Strong Inventories and Self-Directed Search).

Essay, multiple choice, and performance items are some of the cognitive test item types that are used in academic achievement tests. These are often widely classified into objective items and performance assessments. The former are more structured and mostly have only one correct answer. They are divided into two categories: selection-or-recognition-types of items such as multiple-choice, true or false, and matching-type tests, and supply-types items such as sentence completion and short-answer tests.

According to Bandalos (2018), the most versatile of all item test types are the multiple-choice items. It is often concluded that multiple-choice items can only measure information recall and memory sharpness. However, when this type of test is carefully thought and constructed, it is capable of tapping into a much higher level cognitive process like analysis and information synthesis. Test items that require comparison, interpretation of tables and graphs, or creation of new context are examples of item types that require high cognitive reasoning and processes.

Multiple-choice items can also be used to gather diagnostic information regarding a taker's misunderstandings, in addition to cognitive processes (Bandalos, 2018).

1.2 Psychometrics

Psychometrics is the quantitative and technical aspect of measuring mental capabilities. The Psychometric Society was founded in 1935 and it sponsored the journal Psychometrika with its first volume appearing in March 1936. This led to a plea to recognize "a mathematical underpinning for psychological research." Psychometricians, those who are specialists in Psychometrics, are especially keen in providing methods and processes for statistical measurements that can be used widely in psychological research.

According to Appelbaum (1986), the longest-running topic in Psychometrika was perhaps involving computation of the tetrachoric correlation that forms the basis of many approaches in item analysis in test theory.

The study directed on school children by Alfred Binet was the first breakthrough in the study of intelligence. He, then, came up with the Binet scales. This scales and their descendants, together with the IQ concept that is associated with them, continue to be used until today.

Furthermore, David Wechsler and associates extended the intelligence testing to adults and the changed the IQ concept from the mental age system (Mental Age/Chronological Age x 100) to the notion of a deviation IQ that is based on established standards. He was primarily concerned with assessing intelligence of individuals rather than groups. Moreover, as the 20th

Century came, many group-administered paper-and-pencil tests also appeared. These are old Army Alpha and Beta tests, which were created for the screening of inductees in the armed forces during World War I (Goldstein and Hersen, 2000).

In educational, industrial, military, and clinical settings, the psychological or intelligence test became a widely-used assessment instrument. Some tests emphasized gaining an IQ quotient. However, others use them as way to evaluate and measure cognitive processes. (Goldstein and Hersen, 2000).

1.3 Classical Test Theory (CTT)

The Classical Test Theory or CTT is said to be the forerunner in the use of statistics in measuring test scores. It was then called the True Score Theory. It was only distinguished as "classical" eighteen years later in *Statistical Theories of Mental Test Scores,* a book authored by Frederic M. Lord and Melvin Robert Novick and was originally published in 1968.

The CTT has dominated the methods used in the application of test theories to assessments. Charles Spearman figured out how to correct a correlation coefficient due to measurement error and how to solve the reliability index needed in making such correction in 1904. This became the Spearman's model, which was expressed in the following form:

$$X = T + E \qquad \textbf{(1)}$$

Where: X = the observed test score, denoted by ρ^2_{XT};

T = the individual's true score, denoted by σ^2_T;

E = a random error component, denoted by σ^2_X.

Therefore:

$$\rho^2_{XT} = \frac{\sigma^2_T}{\sigma^2_X}$$

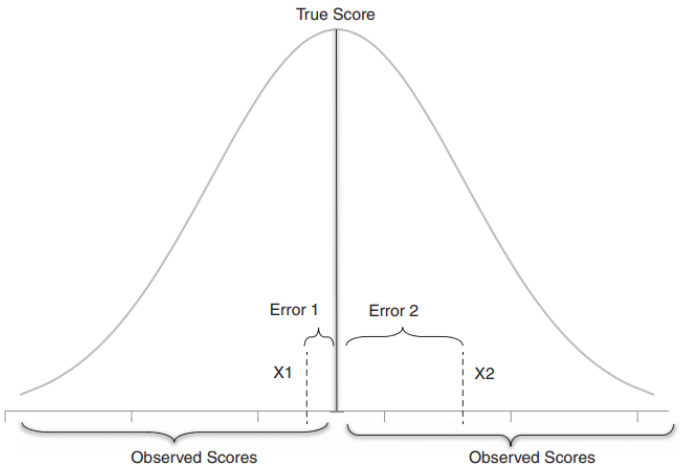

Figure 1. The distribution of observed scores around the true score

Figure 1 shows us the distribution of observed scores around the true score. Moreover, the error scores are seen as being random. If theses error scores were not indeed random, they will have to cancel each other if repeated testing was done. Moreover, the average of these repeated scores would not be equal to the true score. In CTT, the error scores are treated as random and this will result in a normal distribution of observed scores around the true scores (Bandalos, D. L., 2018).

Statistical indices based on CTT has a weak assumption and easier to compute, manipulate and understand; thereby, it is easy to use (Hambleton and Jones, 1933).

1.3.1 CTT Difficulty and Discrimination Indices

Osarumwense & Oyedeji (2015) calculated the item Difficulty Index of an entire number of examinees using the formula:

$$P = R/T \qquad \textbf{(2)}$$

Where: P = item difficulty index,
R = the number of correct responses; and
T = the total number of responses (i.e., correct + incorrect + blank responses)

The computation for the Difficulty Index uses the percentage sample. The scripts were arranged in descending order of the performance of the examinees and the first 27% of the scripts called the upper group U and the last 27% of the scripts called the lower group L were taken the formula:

$$P = \frac{R_U + R_L}{N_U + N_L} \qquad \textbf{(3)}$$

Where: P = Item difficulty index
R_U = the number of examinees who got the item correctly in the upper group,
R_L = the number of examinees in the lower group who got the item correctly,
N_U = Number of examinees of the upper group; and
N_L = number of examinees of the lower group.

For better understanding on the values of the item difficulty index of CTT, the intervals with the corresponding interpretation on Table 1.3.1.1 will be used.

Range	Difficulty Level
0.20 and below	Very difficult
0.21 - 0. 40	Difficult
0.41 - 0.60	Average
0.61 - 0.80	Easy
0.81 and above	Very Easy

Table 1.3.1.1. Interpretation of the Difficulty Index (P)

The Discrimination Index, on the other hand, is computed using the difference between the percentage of students in the upper group (P_U), i.e., the top 27% scorers, who obtained the correct response, and the percentage of those in the lower group (P_L), i.e., the bottom 27% scorers, who obtained the correct response; thus,

$$D = P_U - P_L \quad \textbf{(4)}$$

Where: D = discrimination index
P_U = upper group
P_L= lower group

For better understanding on the values of the item discrimination index of CTT, the intervals with the corresponding interpretation on Table 1.3.1.2 will be used.

Range	Discrimination index
0.40 and above	Very good
0.30 - 0.39	Good item
0.20 - 0.29	Fair item
0.09 - 0.19	Poor item

Table 1.3.1.2. Interpretation of the Discrimination Index (D)

Classical Test Theory approaches are still used today, however, there is also a modern test theory which is known as the Item Response Theory (IRT). CTT has clear shortcomings, thus the reason that modern test theory emerged. IRT was developed to address such issues brought about by CTT.

1.4 Item Response Theory (IRT)

Item Response Theory or IRT is another statistical tool which analyzes the test scores of respondents to each several items or trials are mutually-exclusive categories. IRT is also known as latent trait theory, strong true score theory, or modern mental test theory. It can be applied to a broader and wider scope. In fact, it was developed for purposes of educational assessment and measurement, specifically on student achievement.

IRT has improved immensely the measurement of achievement testing as it overcomes the limitations that was set by CTT. It assumes a continuous latent variable, thus the term ‘latent trait theory,’ that represents the student’s proficiency in responding to test items. The probability of a response in any of two-or-more mutually exclusive categories of an item is assumed to be a function of the location of the student on the latent continuum and of certain estimable parameter characteristic of the item. This process directs to the statistical procedures of test scoring on any number of items without the assumption that these test items are sample from a defined item to which the result generalize (R. Darell Bock and Irini Moustaki, 2007).

In addition, Lee and Cho (2013) stated many e-learning and assessment systems based on IRT are mainly concerned with the ability estimation in order to suggest adjusting learning content or change the test difficulty level in a more customized learning setup. Chang and Yang (2009) also stated that other applications firstly applied IRT for capability estimation and further used classification methods for student rank.

According to Lazarsfeld (1958), item responses being statistically independent, given the respondent's location in latent space, is a further critical assumption in IRT. He made use of the principle of "conditional" independence as an analysis table data.

1.4.1 General IRT Framework

R. Darell Bock and Irini Moustaki (2007) said that the dichotomous, ordered polytomous, nominal polytomous, and ranking are commonly employed modes of response modelled in item response theory.

According to Zheng (2014) multiple choice questions that have dichotomous items, the most common IRT models are the one-parameter logistic (1-PL) model, two-parameter logistic (2-PL) model, and the three-parameter logistic (3-PL) model. The probability of a correct response to item j from an examinee with ability level theta *(θ)* is modeled by the following item response functions (IRFs):

1 –PL; $$P_j(\theta) = \frac{1}{1 + \exp[-(\theta - b_j)]}; \quad \textbf{(5)}$$

2-PL; $$P_j(\theta) = \frac{1}{1 + \exp[-a_j(\theta - b_j)]}; \quad \textbf{(6)}$$

3-PL $$P_j(\theta) = c_j + \frac{1 - c_j}{1 + \exp[-a_j(\theta - b_j)]}. \quad \textbf{(7)}$$

Where:

a_j = parameter of discrimination of item *j,* with $a \in (0, \infty)$,
b_j = parameter of difficulty of item *j,* with $b \in (-\infty, \infty)$,
C_j = parameter of pseudo-guessing of item *j,* with $c \in [0,1]$, and,
θ = level of ability of the examinee, with $\theta \in (-\infty, \infty)$.

Most application of Item Response Theory estimates student's ability basing on two-parameter model (Rasch, G., 1960).

When using IRT method in estimating the ability of a student, Binh and Dui (2016) stated that it depends not only on the number of correct answers but also each item attributes. If two students correctly answer the same item, they must receive the same result. On the other hand, if two students correctly answered the same number of questions but different test items, the result can differ. This makes the two estimation models, CTT and IRT, different from each other. In fact, they can be called linear and nonlinear model, respectively. The one-parameter model sets default for all items with the same difficulty, which is 1. Taking all of those into consideration, these encourage us to use the two-parameters instead of one-parameter.

In estimating ability, according to Baker (2001), there are three methods. These are ability estimation with clear question parameters, question parameters estimation with clear student's ability, and ability and question parameters estimation.

Ability estimation with clear question parameters is the easiest way. The initialized values of an ability will be the beginning of the ability estimation process. The value is, then, employed to calculate the probability of questions with right or correct answers. This value can also be changed further in order to improve the calculated probability value to fit the answered questions result. This process of changing the value will continue until the adjustment value is smaller than threshold value and the estimated ability is not considerably changed. Such process will be done for all the students participating in the test.

1.5 Methods

This study used the descriptive quantitative design, which is a research design that involves observing and describing the behavior of a data (quantitative data) without influencing it in any way. The data used in this study are the raw data from the scored answer sheets of the MSU-TCTO SHSEE given in November 2018, which also served as the research instruments, in analyzing and describing their respective psychometric properties. They were gathered from the Admission Office of MSU TCTO. Stratified sampling was applied in order for the study to avoid biases. The researcher grouped the respondents into different strata according to the municipality in order to have proper distributions of the test takers. Then, the researcher picked in random the envelope of the result from the different municipalities until the desired number of respondents was acquired.

The researcher tallied each correct and wrong answer per test item using the Microsoft Office program, specifically MS Excel, 1 for correct answers and 0 for wrong answers respectively. The name and total scores of the students were represented by numerical values. The study used the formula for the Classical Test Theory (CTT) and Item Response Theory (IRT), specifically the 1PL and 2PL model using the Statistical Program for Social Sciences (SPSS), to determine the difficulty and discrimination index of the said exam. The *t*-Test had been used to determine the significant difference between CTT and IRT. A statistician was consulted for the proper use of the program.

1.6 Results

The following are the results generated using the Statistical Program for Social Sciences (SPSS).

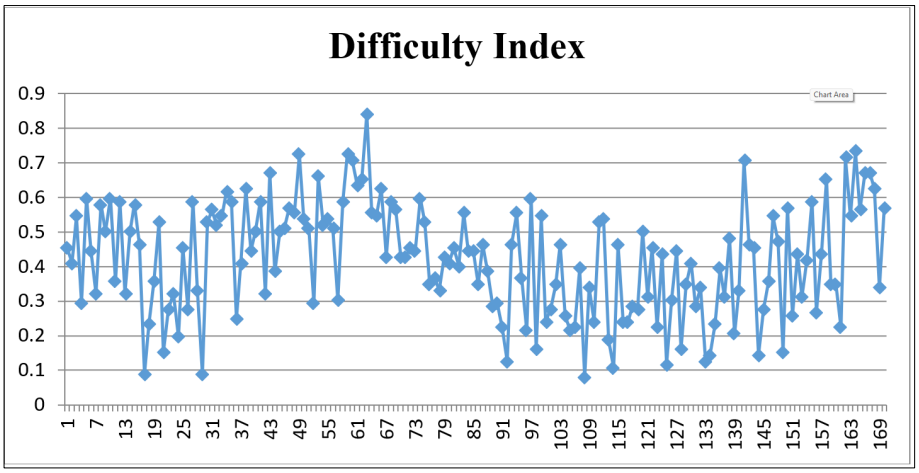

Figure 2: Difficulty Index of the SHSEE using CTT

Results showed that most of the items have difficulty values less than 0.5, which implies that these items are difficult for the takers of MSU-TCTO SHSEE in November 2018. Three of the items that were very difficult are item numbers 17, 29, and 108. On the other hand, only one item is considered to be very easy, which is item number 63. It also showed that most of the items in mathematics and science were very difficult for the test takers while most items in language were moderately easy for them.

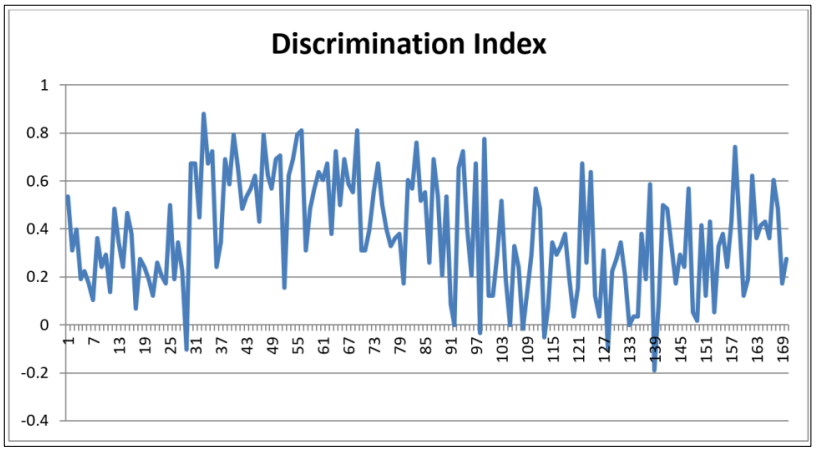

Figure 3: Discrimination Index of the SHSEE using CTT

Using the Classical Test Theory, results showed that there were few items that were below zero discrimination values. This means that these items were poor items and should be subject to removal or revision. Further, most of the items have discrimination values higher than 0.2 which can be considered good items.

Subject	Reliability	Interpretation
Aptitude	0.714	Reliable
Language	0.925	Highly Reliable
Math	0.739	Reliable
Science	0.691	Reliable

Table 2: The Reliability Test for Classical Test Theory by Subject

Moreover, the test in Aptitude, Mathematics, and Science under Classical Test Theory has reliability indices of 0.714, 0.739, and 0.691 respectively which are interpreted as reliable. While the test in Language with a reliability index of 0.925 is interpreted as highly reliable.

Subject	Reliability	Interpretation
Aptitude	0.974	Highly Reliable
Language	0.965	Highly Reliable
Math	0.967	Highly Reliable
Science	0.983	Highly Reliable

Table 3: The Reliability Test for Item Response Theory

Under the Item Response Theory, the Aptitude category got reliability of 0.974 meaning it is highly reliable. Language has a reliability score of 0.965 the interpretation it is also highly reliable. Math and Science got a reliability score of 0.967 and 0.983 respectively meaning they're also highly reliable. The tendency for IRT to have higher reliability is due to the approximation of the true variance since the data collected is on part of the population.

Test Theory	Overall Reliability	Interpretation
Classical Test Theory	0.939	Highly Reliable
Item Response Theory	0.968	Highly Reliable

Table 4: Overall Reliability of the Item under CTT and IRT

t-value	*p*-value	Interpretation
-3.679	0.035	No significant difference

Table 5: Comparison between IRT and CTT for Reliability

Both CTT and IRT were highly reliable with a reliability index of 0.939 and 0.968. By comparison, the IRT had a slightly higher reliability index than the CTT. The difference might be due to the definition of the true variance in IRT that the distribution was normally distributed with mean zero and variance one. The t-value was -3.679 with a p-value of 0.035 which is less than the level of significance of 0.05, this means that there is a significant difference between the IRT model and the CTT approach. However, is not significant at a 0.01 level of significance. The result implies that the difference was about 95% level of confidence only.

The following are the findings of the study:

1.) the result of the SHSEE under CTT for Language is highly reliable, meaning highly adequate and acceptable;
2.) the results of the SHSEE under CTT for Aptitude, Mathematics and Science are reliable, meaning adequate and acceptable;
3.) the overall result of the SHSEE under CTT is highly reliable, meaning highly adequate and acceptable;
4.) the results of the SHSEE under IRT for Aptitude, Language, Mathematics, and Science are highly reliable, meaning highly adequate and acceptable;
5.) the overall result of the SHSEE under IRT is highly reliable, meaning highly adequate and acceptable; and,
6.) the results showed that the examination both have high reliability, meaning high adequacy, under both of the Classical Test Theory and Item Response Theory.

Conclusion

Based on the results and findings, the following conclusions are obtained in this study. The test items of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Examination were highly adequate and reliable both CTT and IRT. Furthermore, there is a significant difference between the reliability index under IRT and CTT models at 0.05 level of significance, but not at 0.01 level of significance. Therefore, there is slight inconsistency of the result. Some of the items need to be revised in order to come up with reasonable passers for SHSEE.

This informs that the MSU-TCTO Senior High Administration and SHSEE Steering Committee shall continue to enhance the entrance examination for the next batches, the MSU TCTO SHSEE committee may use Item Response Theory rather than Classical Test Theory as statistical treatment since it gives more emphasis on each item in the assessment of the reliability of the questions in the examination.

Acknowledgment

This study is a thesis requirement for the researcher's Masters's degree in Science in Teaching Mathematics at the Graduate School of Mindanao State University – Tawi-Tawi College of Technology and Oceanography. This was conducted from November 2019 to January 2020.

My heartfelt gratitude to Mr. Ummar A. Sallil, MSc for his valuable insights and for serving as the statistician of this research, and to Mr. Ladznar S. Laja, PhD, Chief of the Admissions Office, and Ms. Anabel A. Wellms, PhD, Chairperson of the Senior High School Entrance Examination Steering Committee, of the Mindanao State University - Tawi-Tawi College of Technology and Oceanography for the permission to use the results of the November 2018 examination as data to this study.

References

Awopeju, O. A., (2008). *Comparative Analysis of Classical Test Theory and Item Response Theory Based Item Parameter Estimates of Senior School Certificate Mathematics Examination*. doi: 10.19044/esj.2016.v12n28p263

Baker, F. B., (2001). *The Basic of Item Response Theory*. ERIC Clearing house on Assessment and Evaluation. University of Wisconsin, 2001

Bandalos, D. L., (2018). *Measurement Theory and Application for the Social Sciences. The Guildform Press New York London*. pp 63-69, 120, 157, 159, 404, 407, 420.

Bock, D. and Moustaki, I. (2007). *Item response theory in a general Framework in Handbook of Statistics on Psychometrics, Vol. 26, edited by C. R. Rao and S. Sinharay. Elsevier.*

Bridgeman, B., and Cline, F. (2000). *Variations in Mean Response Times for Questions on the Computer-Adaptive GRE General Test: Implications for Fair Assessment (ETS RR-00-7).* Available online at: https://www.ets.org/research/policy_research_reports/publications/report/2000/hsdr

Chang, W., Yang, H., (2009). Applying IRT to estimate learning ability and k-means clustering in web based learning. *Journal of Software.* Vol.4, No.2.

Divgi, P. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58, 357-381.

Goldstein G, Hersen M. (2000). *Handbook of psychological assessment (3rd ed.).* Oxford, United Kingdom.

Gullicksen H. (1950). *Theories of Mental Test Score*. New York.

Hambleton, R. K., Jones, R. W.. *Comparison of Classical Test Theory and Item Response Theory and their Application to Test Development*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1690.7561&rep1&type=pdf

Hoang Tieu, Binh & Bui, The. (2016). *Student ability estimation based on IRT.* 10.1109/NICS.2016.7725667.

Lawson, S. (1991). *One Parameter latent trait measurement. Do the results justify the effort?* In B. Thompson (Ed.), Advance in education research: Substantive findings, methodological development, 1, 159-168.

Lazarsfeld, P. F. (1958). *Evidence and inference in social research*, Dedalus, 87, 99-109.

Lee, Y., & Cho, J. (2013). Personalized item generation method for adaptive testing systems. *Multimed Tools Appl,* 74(19): 8571-8591.

Maloney M.P., & Ward M.P. (1976). *Psychological Assessment. A Conceptual Approach.* New York: Oxford University Press.

Ojerinde (2013). *Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of Comparability of item Analysis Results.* Lecture Presentation at the Institute of Education, University of Ibadan.

Osarumwense, H. J., Oyedeji, S. O. (2015). *Empirical Comparison of Methods of Establishing Item Difficulty Index of Test Items Using Classical Test Theory (CTT).*

Paul, M., & Sampo, V. (2002). *A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory*. Educational and Psychological Measurement, 6, 921-943.

Rasch, G. (1960). *Probabilistic model for some intelligence and attaintment tests*. Copenhagen: Danish Institute for Educational Research.

Royce, H. (2009). Comparison of the item discrimination and item difficulty of the quick-mental aptitude test using CTT an IRT methods. The international Journal of Education and Psychological Assessment. Vol. 1, Issue 1, pp. 12-18.

Sallil, U., 2017. *Estimating Examinee's Ability in a Computerized Adaptive Testing and Non-Adaptive Testing using 3 parameters IRT model.*

Spearman, C. (1907). *Demonstration of formulae for true measurement of correlation. American Journal of Psychology,* 18, 161-169.

Stone, C. A., Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models Using SAS.* SAS Institute Inc., Cary, NC, USA.

Zheng, Y. (2014). *New Methods of Online Calibration for Item Bank Replenishment*