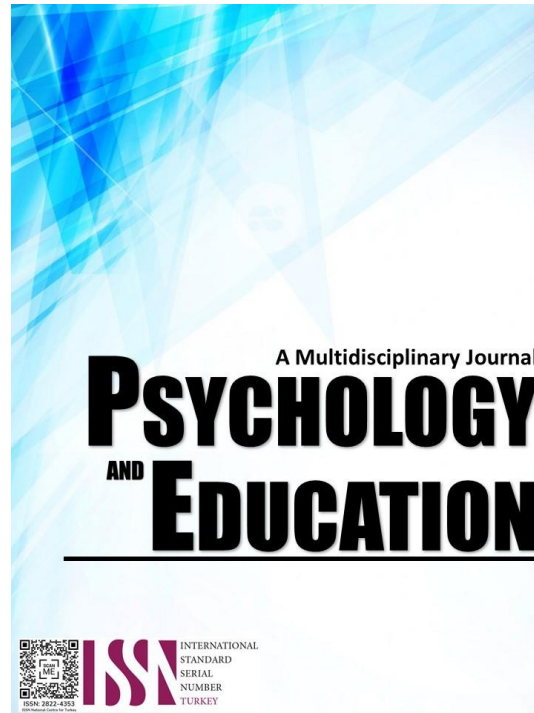


EVALUATING THE RELIABILITY INDEX OF AN ENTRANCE EXAM USING ITEM RESPONSE THEORY



PSYCHOLOGY AND EDUCATION: A MULTIDISCIPLINARY JOURNAL

2023

Volume: 7

Pages: 656-658

Document ID: 2023PEMJ587

DOI: 10.5281/zenodo.7750347

Manuscript Accepted: 2023-15-3

Evaluating the Reliability Index of an Entrance Exam Using Item Response Theory

Jeffrey Imer C. Salim*

For affiliations and correspondence, see the last page.

Abstract

Humans, in the conduct of daily activities, have used measurement as a crucial tool. In education, certain measurement tools such as achievement testing is used to assess the psychological capabilities of students. Thus, it is important that correct test constructs are utilized for results to be reliable. This study was done to evaluate the reliability index of the Mindanao State University – Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Examination, which is given annually to prospect students. The examination consists of four subjects: English, Science, Mathematics and Aptitude with 75, 30, 40, and 25 questions each respectively. The study employed the descriptive quantitative research design. Stratified sampling was used on the entire raw scored answer sheets to come up with a sample of 200 examinees of the said exam. To evaluate the reliability index, the Statistical Program from Social Sciences or SPSS was used. The study concluded that the test constructs for the examination was highly reliable with reliability index of 0.968, with a level of significance of 0.1. Reliability per subject were 0.965 for Language or English, 0.983 for Science, 0.967 for Mathematics, and 0.974 for Aptitude. The study recommends that the examination be further evaluated and updated.

Keywords: *measurement, achievement testing, reliability index, item response theory*

Introduction

The primary objective of psychological measurement is to distinguish the psychological attributes of individuals and the differences among them. Moreover, measurement theory is a branch of applied statistics that describes and evaluates the quality of measurements (including the responses process that generates specific score patterns by persons), with the objective of improving their usefulness and accuracy (Price, 2017). Psychometricians use measurement theory to propose and evaluate methods for developing new tests and other measurement instruments.

According to Gregory (2000), psychological measures or assessment measures involve quantification techniques. In other words, the psychological measure is a vibrant process, repeatedly changing, but retaining its original structures. Moreover, a psychological measurement can be a complicated process with the following characteristics: scores or categories, behavior samples, norms and standards, standardized procedures, and prediction of non-test behavior.

Literature Review

Measurement and Achievement Testing

The problem of improving and quantifying the psychological measurement is addressed by doing a

psychological testing. 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 cognitive items that are used in academic achievement tests. These are often widely categorized into objective items and performance assessments. The most versatile of all item types are the multiple-choice items. It is often concluded that multiple-choice items can only measure rote recall of information, when they are cleverly constructed, can tap into higher-level cognitive process such as analysis and synthesis of information. Items that require to detect respondent similarities or differences, interpret graphs or tables, make comparison, or mold previously learned material into a new context emphasizes on higher-level cognitive processes. And these are appropriate for wide variety of subject matter. Another benefit of multiple-choice items is the fact that they can provide useful diagnostic information regarding the respondent's misunderstandings. Fails or distractors or incorrect options must be based on common misconceptions or errors (Bandalos, D.L., 2018).

A term that refers to a wide range of strategies, both

qualitative and quantitative, that is used to assess the quality of pool items is called item analysis. When measuring behavioral or psychological data, it is called psychological scaling. This is due to the simple fact that the assignment of numerals places the objects or events on a scale. And to make proper use of the behavioral or psychological data, the assignment of numerals is mostly based on mathematical or statistical models for those data (Jones and Thissen, 2007).

Item Response Theory

Item Response Theory or IRT deals with the statistical analysis of data wherein responses of each of a number of respondents to each of a number of items or trials are mutually-exclusive categories. IRT has a broader and wider potential application. It was, however, developed primarily for educational measurement, specifically the measurement of an individual student achievement. Before IRT, the treatment of achievement test data statistically was based entirely on the classical test theory.

In 2016, Binh and Dui stated that IRT can also be coined hidden latent model because the latent is discovered by observing the variables and properties or parameters of the models. Moreover, many applications using IRT have concentrated on the estimation of student's ability to give suggestions to adapt learner's tutorial or documentation in recent years.

Lee and Cho (2013) stated many e-learning and assessment systems based on IRT are mainly concerned with ability estimation in order to suggest adjusting learning content or change test difficulty level in personalized learning systems. In addition, Chang and Yang (2009) stated that other applications firstly applied IRT for cabality 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.

Methodology

Descriptive quantitative design involving observing and describing the behavior of quantitative data sans influencing them was used for the conduct of this study. The raw data are from the scored answer sheets of the Senior High School Entrance Examination of Mindanao State University – Tawi-Tawi College of

Technology and Oceanography, which was given on November 2018. To avoid subjective biases, stratified sampling where respondents are stratified per municipality was done to have proper distribution. Then, a random envelope containing an answer sheet was picked until respondents' number were completed.

Each answer was tallied using Microsoft Excel, 1 for correct answers, and 0 for wrong ones. Moreover, the name and total scores of the students were changed to numerical values. The study used the formula for the Item Response Theory (IRT) using the Statistical Program for Social Sciences (SPSS), to determine the reliability level by subject and its totality. A statistician was also consulted for the proper computations.

Results and Discussion

The following are the results from the SPSS:

Table 1. *Reliability Indices by Test Subjects*

<i>Subject</i>	<i>Reliability</i>	<i>Interpretation</i>
Aptitude	0.974	Highly Reliable
Language	0.965	Highly Reliable
Math	0.967	Highly Reliable
Science	0.983	Highly Reliable

Table 2. *Reliability Index of the Examination*

<i>Test Theory</i>	<i>Overall Reliability</i>	<i>Interpretation</i>
Item Response Theory	0.968	Highly Reliable

Table 1 shows the reliability indices with corresponding interpretation by subjects using the Item Response Theory. Results showed that the Science subject has the highest reliability index of 0.983. Aptitude came next with 0.974. Language and Mathematics have similar reliability indices of 0.967. This means that the test items in these subjects are highly reliable. The examination garnered a reliability index of 0.968 using the Item Response Theory, which means that the examination is highly reliable in assessing the examinees.

Conclusion

This study assessed the reliability index of the test items by subject and as a whole of the Mindanao State University-Tawi-Tawi College of Technology and Oceanography Senior High School Entrance Examination conducted on October 2018 under Item Response Theory.

Data were collected with the approval of the Dean of the MSU TCTO College of Arts and Sciences, who is also the chairperson of the committee of such examination, and the Admissions Office. The answer sheets of the takers of the said examination served as the raw data and were subjected to a statistical treatment, Item Response Theory, to find out the reliability index of the test items by subject and its totality.

Based on the results and findings, the following conclusions are obtained in this study: (1) The results of the examination under IRT for Aptitude, Language, Mathematics, and Science are highly reliable, meaning highly adequate and acceptable. (2) The overall result of the examination under IRT is highly reliable, meaning highly adequate and acceptable. (3) The study recommends that the examination committee further enhances the examination and use Item Response Theory in the evaluation of its reliability.

References

- Aldrich, J. (1997). "R.A. Fisher and the making of maximum likelihood 1912-1922." *Statist. Sci.* 12 (3) 162 - 176, August 1997. <https://doi.org/10.1214/ss/1030037906>
- Baker, F. B., (2001). *The Basic of Item Response Theory*. ERIC Clearing house on Assessment and Evaluation. University of Wisconsin,
- 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.
- Binh, H.T. & Dui, B.T. (2016). *Student ability estimation based on IRT*
- 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
- 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.
- Gregory, R.J. (2000). *Psychological Testing*. 3rd Edition. Illinois: Allyn and Bacon, Inc.
- Hassan, S. & Hod, R. (2017). Use of Item Analysis to Improve the Quality of Single Best Answer Multiple Choice Question in Summative Assessment of Undergraduate Medical Students in Malaysia. *Education in Medicine Journal*. 2017;9(3):33-43. <https://doi.org/10.21315/eimj2017.9.3.4>
- Jones, Lyle & Thissen, David. (2006). 1 A History and Overview of Psychometrics. *Handbook of Statistics*. 26. 1-27. 10.1016/S0169-7161(06)26001-2.
- Lee, Y., & Cho, J. (2013). Personalized item generation method for adaptive testing systems. *Multimed Tools Appl*, 74(19): 8571-8591.
- Lazarsfeld, P. F. (1958). Evidence and inference in social research, *Dedalus*, 87, 99-109.
- Price, L. R. (2017). *Theory into Practice*. The Guild press New York London. pp. 5
- Rasch, G. (1960). *Probabilistic model for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Sallil, U., 2017. Estimating Examinee's Ability in a Computerized Adaptive Testing and Non-Adaptive Testing using 3 parameters IRT model.
- Warm, T.A. Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427-450 (1989). <https://doi.org/10.1007/BF02294627>
- Zheng, Y. (2014). *New Methods of Online Calibration for Item Bank Replenishment*

Affiliations and Corresponding Informations

Jeffrey Imer C. Salim

Mindanao State University -Tawi-Tawi College of Technology and Oceanography - Philippines