Systematic Differences in Item Difficulty Analysis between Classical Test Theory and Item Response Theory

Maria Nelly Gutierrez Arvizu

Northern Arizona University

Abstract

A study comparing item analysis under the scope of Classical Test Theory (CTT) and Item Response Theory (IRT) was conducted using 770 responses to eight placement tests in an Intensive English Program. The purpose of the study was to investigate the systematic differences found when comparing item difficulty, standard error of the mean, and reliability in both approaches to item analysis. The data was handled by testlets (items from the same reading or listening passage). There was a total of 83 testlets (49 listening and 34 reading). The data was analyzed using Excel and SPSS (for CTT) and FACETS (for IRT). The results showed that the item difficulty level (easy, medium, and difficult) was different in CTT and IRT. IRT had more items classified as medium as it accounts for examinees' proficiency level. The standard error of the mean was found to be similar in both frameworks. Finally, the difference in reliability is in the definition of the reliability concept in CTT and IRT.

Systematic Differences in Item Difficulty Analysis between Classical Test Theory and Item Response Theory

Background

Item analysis is used for a variety of purposes in testing: 1. to determine someone's ability using a score; 2. to select the most suitable items for a test; 3. to admit someone in a program; 4. to place someone according to his/her level of proficiency. There are two approaches to analyzing how items behave: Classical Test Theory (CTT) and Item Response Theory (IRT). CTT uses statistics such item difficulty (proportion of correct responses), standard error of the mean, and Cronbach's Alpha to determine the ability of a group of examinees (Hambleton & Jones, 1993). Item Response Theory (IRT) is based on the one-parameter Rasch logistic model, in which persons' abilities and item difficulties are placed in the same scale (Bond & Fox, 2001; Myford, 2006). As educators or researchers, it is important to understand the information these approaches provide and our purpose in testing in order to determine which framework is the most suitable.

Research Questions

There are three research questions driving this study.

1. What are the systematic differences in item difficulty between CTT and IRT when assessing listening and reading skills using dichotomous in item difficulty?
2. What are the systematic differences in item difficulty between CTT and IRT when assessing listening and reading skills using dichotomous in standard error of the mean?

3. What are the systematic differences in item difficulty between CTT and IRT when assessing listening and reading skills using dichotomous in reliability?

Methods

A total sample of 770 responses to eight placement tests administered from Fall 2009 to Spring 2013 were analyzed. These tests had a total of 83 testlets (49 Listening, 34 Reading) with 541 items. The testlets were used for the comparison. In CTT, the item difficulty coefficients were averaged to obtain an item difficulty coefficient per testlet. In IRT, item difficulty per testlet was obtained from FACETS as 'testlet' was used as one of the facets in the model.

Item difficulty, standard error of the mean, and reliability were computed in CTT and IRT. Excel and SPSS were used for CCT and FACETS was used for IRT. All of the coefficients for item difficulty were classified in easy, medium, or hard for comparison purposes. Item difficulty and standard error of the mean were compared to identify if they were similar or different.

Results

Results showed that there are differences among the approaches particularly in item level of difficulty. Thirty three (out of 83) testlets had a difference in item difficulty level. Moreover, most of the testlets in IRT were considered of medium difficulty, 70 in IRT compared to 48 in CTT. It is important to consider that CTT is sample dependent while IRT accounts for differences in examinees' abilities. This could be the source of the difference.

The standard error of the mean was very similar in both approaches. Even though the coefficients might appear to be very different, IRT is on a scale from -1 to 1 while CTT is on a

scale of 0 to 1. The criterion used to determine if they were different was a difference of .03 between the two coefficients, as this would exceed the 95% Confidence Interval by ±.10. Only six (out of 83) testlets were different when comparing the standard error of the mean.

Reliability was measured by test divided in two sections (listening and reading). Reliability was computed in a total of 16 sections (8 listening and 8 reading). The reliability coefficients in this study were very similar (14 similar out of 16); however, they measure different aspects. CTT uses Cronbach's alpha to measure internal consistency, that is, how well items in a test work together. IRT provides a reliability coefficient to measure how well items are differentiated. Even though the term reliability is used in both approaches, CTT and IRT measure different aspects and researchers should be careful when interpreting these.

Relevance to PIE

The results presented in this study might be used in the development of a computer adaptive placement test at the PIE. This would allow for a more efficient administration of the placement test. The number of items needed to place an examinee into their level could be reduced, thus the time it takes to complete the test and provide results.

In addition, it can serve the purpose of selecting, deleting, or modifying items in future administrations of the placement test. The data gathered for this project could be used in future research projects to compare the results of the same item or testlet at different administrations. A historical analysis of the placement tests can be conducted, as well.

References

Bond,T., & Fox, C. (2001). *Applying the Rasch Model. Mahwah, NJ: Lawrence Erlbaum Associates.*

Hambleton, R. K., & Jones, R.W. (1993). Comparison of Classical Test Theory and their applications to test development. *ITEMS, Fall,* 38-47.

Myford, C. (2006). Analyzing rating data using Linacre's FACETS computer program: A set of training materials to learn to run the program and interpret output.