International conference
The Future of Education
2nd Edition
PIXEL

Test Development for Grade Eight Students: A paradigm shift from Classical Test Theory to Item Response Theory

Indrani Bhaduri
National Council of Educational Research and Training (India)
indranibhaduri@gmail.com

Introduction

The worth of any educational assessment endeavour depends on the instruments i.e. the tools and techniques used, if these instruments are poorly designed, the assessment can be a waste of time and money. The same is true while conducting large scale achievement surveys. The precursor to conducting the achievement survey is the development of a robust instrument, thus test development is of prime concern while conducting surveys. In India, the Ministry of Human Resource and Development (MHRD), has designated the National Council of Educational Research and Training (NCERT) as the national authority for conducting the national achievement surveys. The present survey conducted in 2012, is the third cycle of Grade VIII in which the children's learning achievement has been measured in Sciences, Social Sciences, Mathematics and Language. This paper analyzes the process of test development in Science and is based on the data collected during the field try out of the tests administered to 7638 students from 11 States and Union Territories of the country. In this survey student responses to questions in the tests were analysed using modern Item Response Theory (IRT) rather than the classical techniques. This practice of using IRT, is used in the major international surveys such as the Programme for International Student Assessment (PISA), Progress in International Reading Literacy Study (PIRLS) and Trends in International Mathematics and Science Studies (TIMSS). The present study probed into the consequences of using a complex test and item analysis approach in a situation that historically has used a simple conventional approach. It queried as to whether the increase in complexity and difficulty associated with the use of IRT pay significant dividends in improving the tests?

Theoretical Framework

Classical test theory (CTT) and Item response theory (IRT) embodies two different measurement frameworks. Although CTT has been in vogue for most of this century, IRT has observed a phenomenal growth in recent decades. Though CTT is easy to apply in many testing situations (Hambleton & Jones, 1993) as it has relatively weak theoretical assumptions which are easy to meet, the major limitation of CTT is that the person statistic (i.e., observed score) is item dependent, and the item statistics (i.e., item difficulty and item discrimination) are examinee dependent. This circular dependency poses a major difficulty in educational measurements. IRT, on the other hand, models the probabilistic distribution of examinees' success at the item level. IRT primarily focuses on the item-level information in contrast to the CTT's primary focus on test-level information. The IRT framework includes a group of models, and the applicability of each model in a particular situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, there are three IRT models, known as three-, two-, and
one-parameter IRT models. Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. As a result, in theory, IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991).

Methods

The third cycle of the National Achievement Survey for Grade VIII was initiated in 2011 with a view to measure the children's learning achievement in Sciences, Social Sciences, Mathematics and Language. The process of test development was commenced with the preparation of an assessment framework. The framework developed provided an understanding of the construct being assessed and also the processes associated with the construct. It included the definition of the construct and the task for collecting evidences for the understanding of that construct by the students. The Science assessment framework was organized around two dimensions, a content dimension specifying the domains or subject matter to be assessed within science (biology, physics and chemistry) and cognitive dimension specifying the thinking processes to be assessed (that is knowing, applying and reasoning). The cognitive domains described the sets of abilities expected of students as they engaged with these science content tasks. After listing out the topics to be covered in each domain and delineating the claims and corresponding evidences, the tasks for collecting the evidences were developed. In all 180 items were developed and were distributed in three forms. Each form consisted of 20 anchor item and 40 unique items. The total number of items in each form was 60. The items were categorized into three levels of difficulty i.e. easy, average and difficult. In each form, 30% of the items were easy, 45% average and 25% of the items were judged to be as difficult. The piloting of the items were done in 11 states, namely Goa, Sikkim, Nagaland, Meghalaya, Mizoram, Bihar, Uttar Pradesh, Rajasthan, Haryana, Himachal Pradesh, and Madhya Pradesh and in two languages i.e. in Hindi and in English. Analysis of the piloted items was done using classical analysis as well as item response theory. Based on the statistical analysis and the review of pool of items by subject's experts and technical experts a total of 120 items were finally selected. Two test forms were developed with 60 items each wherein item numbers 1 to 15 were unique to each test form, 16 to 45 were the anchor items or the common items in both the forms and 46 to 60 were again unique to both the forms.

Results

The results obtained after the classical analysis for all the 180 piloted items shows that there is a difference in the difficulty level of the items as predicted by the experts, before administration of the test compared to the difficulty level found empirically after the test administration. The judged difficulty level of the items were categorized into three levels of difficulty by the experts i.e. easy (E), average (A) and difficult (D). The items were judged to be as easy if it could be correctly done by 60% or more of the students in the sampled group, it was considered as of average level of difficulty if 30% to less than 60% students sampled could do it correctly and it was difficult if less than 30% were able to it correctly. Out of the 180 items prepared, the experts had opined that in both the languages, 64 items were easy, 75 items were of average difficulty and 41 were difficult. The analysis of the data after the field administration showed that in English medium 5 items were easy, 101 were average and 41 difficult and in the Hindi translation 2 were easy, 84 average and 94 difficult. Analysis was also done in terms of differential item functioning (DIF) of each item following the classification given by Educational Testing Service (ETS), Princeton, US., which says, an item with a DIF value of “C”, should be rejected, getting a value of “B” the item has average DIF and can be accepted with modification, and an “A” there is no DIF and the item can be accepted. It was found that 132 items were categorized as “A”, 25 as “B” and 23 as “C”.

The item response theory analysis was done by plotting the individual Item Characteristic Curves (ICC) for each item. It provided the item wise analysis rather than the test as a whole. Also the concept of item parameter is fundamental to IRT. The IRT models are based on one, two, or three parameters. In this study, the 2 parameter logistic model (2PL) was used. The ‘a’ and the ‘b’ values were similar to the ones obtained for discrimination and difficulty values for the CTT. The *‘a’* parameter expressed how well an item can differentiate among examinees with different ability levels. Study of literature shows that the good items usually have discrimination values ranging from 0.5 to 2.00. As mentioned in the methodology the items were piloted in two different languages Hindi and English and the results of the discrimination values of the items in the mediums English and Hindi ranged from 0.19 to 0.65 and 0.12 to 0.69 respectively. The difficulty of an item, the *b* parameter is the point where the ICC has the steepest slope. The more difficult an item is, the higher an examinee’s ability must be in order to answer the item correctly. Items with high *b* values are hard items, which low-ability examinees are unlikely to answer correctly. Items with low *b* values are easy items, which most examinees, including those with low ability, will have at least a moderate chance of answering correctly. The analysis of the ICC showed that 41% of the items in English and 52% of the items in Hindi had high *b* values. In addition to the plotting of the ICC curves, the IRT was useful to determine the effect of adding or deleting a given test items or a set of test items on the Test Information Function (TIF) and the Standard Error (TSE) function. The TIF is the sum of the item information functions for the items being examined. After adding or deleting items and then examining the change in the shape of the TIF and TSE functions, and comparing it to the desired performance curve, the test was tailored closely to the required specifications. Another advantage of using IRT in Test development is Matrix Sampling. Testing at the National level plays an important part in evaluating the effectiveness of a particular scheme being operated at the National level in the schools. In this the main concern is to identify the strengths and weaknesses of the scheme in promoting student achievement in the various curriculum in operation across the different states in the country. It is impractical to test the sampled students in each and every domain of the curriculum. IRT helps in minimizing the burden on the students by employing matrix sampling method, in which the sampled students responds only to one of the test forms developed which contains a small number of the items representing the category.

While developing the final test forms all the above discussed constructs were considered and finally out of the total of 180 items piloted, 90 items were selected and two test forms were prepared.

Discussion

For developing tests for the national assessment of educational achievement the following steps were followed:

- Developing an assessment framework
- Item writing
- Expert panel review (Content and Bias review)
- Field test of the items
- Statistical review
- Selecting the test items
- Producing the final test

In the present assessment, IRT was used extensively in the process of test development. IRT assumes that test-item responses by students are the result of underlying levels of knowledge and skills, known as ability, possessed by those students. Items that fit the IRT model have lower probabilities of correct responses from low-achieving students and higher probabilities of correct responses from high-achieving students. This was reflected in the item characteristic curve (ICC) of the items. Two important functions, derived from IRT parameters, were used to describe how well the test is functioning: the Test Characteristic Function, which represents the average of all ICCs on the test, and the Test Information Function, which reflects the test’s reliability by providing overall test precision information. Both functions play a critical role in test development and evaluation. These functions were used for the present test construction and selection of test items.IRT therefore helped to produce a test that had the desired precision of measurement at the defined ability level. With respect to test scoring too the IRT-based methods are more useful compared to CTT, i.e. they offered considerable advantages over the “number right” scoring methods typically used in CTT-based tests. For example, when estimating an examinee’s score using IRT, it simultaneously consider the following sources of information i.e. which items were answered correctly/incorrectly and for each of those items, the difficulty and discrimination of the item. This information has the potential to produce better estimates of the ability scores, to produce quantitative estimates of the “quality” or likelihood of any given observed response profile and to assess the degree to which the given IRT model provides a good “fit” to the pattern of responses produced by the individual in question.The use of IRT thus, helped to link the probabilities of item responses to the characteristics assessed by the test. It was also useful to assess the test’s measurement accuracy when individual items were added or deleted in terms of Test Information Function and the Standard Error Function. Also, analyzing the test at the level of individual items facilitated to address problems beyond the scope of CTT. An additional gain of using IRT is matrix sampling. Thus, IRT method is indeed remarkable as it offers a powerful and flexible method for test development, scoring, and evaluation, they represent a vast improvement over approaches based on classical test theory.

References

[1] De Ayala, R. J. (2009). The theory and practice of item response theory. New York: NY: Guilford.

[2] Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. New York, NY: Psychology Press.

[3] Hambleton, R. K., Swaminathan, H., and H. J.Rogers (1991). Fundamentals of item response theory. Newbury Park, CA: SAGE.

[4] Schnakenberg, Keith E. and Fariss, Christopher J., A Dynamic Ordinal Item Response Theory Model with Application to Human Rights Data (January 9, 2010). APSA 2011 Annual Meeting Paper. Available at SSRN: http://ssrn.com/abstract=1534335 or http://dx.doi.org/10.2139/ssrn.1534335

[5] Ying Lu, "Assessing fit of item response theory models" (January1,2006). Electronic Doctoral Dissertations for UMass Amherst. Paper AAI3206198. Available at http://scholarworks.umass.edu/dissertations/AAI3206198

[6] Zimowski, M.F., Muraki, E., Mislevy, R.J., and Bock, R.D. (2000). BILOG-3: IRT Analysis and Test Maintenance for Binary Items, Chicago: Scientific Software, Inc.