The International Journal of Educational and Psychological Assessment
December 2009, Vol. 3

Two Theories, One Theta: A Gentle Introduction to Item Response Theory as an Alternative to Classical Test Theory

Rajiv Amarnani
De La Salle University, Manila, Philippines

Abstract

Item response theory is a relatively recent alternative to classical test theory in psychometrics. Classical test theory concerns itself with determining the observed score, standard error of measurement, true score in a test, then comparing the observed score with norms or a preset criterion in order to interpret that score. In classical test theory, all items are created equal and contribute equally to the obtained score. In contrast, item response theory concerns itself with the information value it can derive from each item, maximizing input from item difficulty, item discrimination, and the possibility of guessing to produce a maximum likelihood estimation of theta, a symbol representing the psychological phenomenon of interest. The various models and assumptions of item response theory are discussed and explored. As an instructive example, another study that compared analyses performed in classical test theory and item response theory will also be discussed.

Suppose you are in the market for a new cellular phone. These days, cellphones boast all sorts of features, ranging from built-in video cameras to GPS tracking systems. The sheer variety of features in cellphones today can make picking a specific model surprisingly difficult. Does one go for the cellphone with a huge capacity for storing music or the one with a five-megapixel camera? How about the cellphone with both these features but double the price as well?

One may choose to list the pros and cons of each cellphone you're interested in and weigh them in against each other, eventually picking the one with the highest pro-to-con ratio. This approach treats each feature of the putative cellphones as equivalent in value. However, a more common approach (and one that is far more intuitive), would be to weight each feature of the cellphone according to its value to you. Are you more of a music lover, or are you more of a budding photographer? Naturally, not all of the features that your cellphone throws at you are of equal value, and one is likely to make decisions and evaluations based largely on the features that are most "appealing" to you or the features that best fit your needs.

These approaches to determining the value of a cellular phone are analogous to the approaches used in assessing one's abilities, opinions, or feelings—the very crux of psychometrics. The first approach, much like classical test theory (CTT), essentially treats each test (or cellphone) as a generally discrete, atomic whole. Each item, regardless of difficulty or predictive power, contributes equally to the raw score one obtains in the test, and, naturally, to the scaled scores and eventually to the assessment itself (Lord, 1977).

Item response theory (IRT), a more recently developed approach to psychometric theory, sees things a fair bit differently. In IRT, the difficulty and the predictive power of each item are factored into the obtained score (Hambleton, Swaminathan, & Rogers, 1991). This takes into consideration the relative value of

 ISSN 2094-0734

each item (or feature) in determining theta (the individual's level of the psychological property measured by the test). The underlying theory simply runs in converse: For every possible theta (or psychological property level), there exists a weighted score that corresponds to that level of psychological ability, opinion, or feeling (Hambleton, Swaminathan, & Rogers, 1991).

The primary difference between CTT and IRT is in the information that each employs in determining an individual's true score (Anastasi & Urbina, 2005). CTT takes a safer route, essentially compiling the psychometric power and the standard error of each item to produce a robust test that can withstand the subversions of other individual differences that may otherwise adversely affect the eventual score in that test (Sijtsima & Bunker, 2006). On the other hand, IRT takes a much more daring approach to psychometrics by incorporating more information about each item in order to produce a clearer and more accurate depiction of each individual's theta (Fan, 1998).

CTT and IRT are two psychometric cameras that take slightly different pictures of an individual's theta. CTT is your 0.5 megapixel camera that can take a decent picture of one's theta without substantial energy consumption (battery power), whereas IRT is your 5 megapixel camera that takes a far more crisp picture of one's theta. In IRT, however, the increased resolution comes at the cost of substantial energy consumption (i.e. you need more batteries). The batteries that power the camera are no different from the psychometric properties of the test items that power the assumptions beneath each psychometric theory. IRT requires far more computational rigor, but the payoffs are arguably commensurate to the cost.

One of the major drawbacks of CTT is that the score obtained (the observed score) is simply an estimation of the true score, or theta (Hambleton, 2000). Another is that, since each test has its own level of difficulty, an individual's scores on different tests are not readily comparable (Hambleton, 2000). For example, a student that has a higher score in Exam 1 than in Exam 2 did not necessarily do better in Exam 1—perhaps Exam 2 was just more difficult than Exam 1. Furthermore, all norms obtained would undoubtedly depend on the normative sample (Fan, 1998). Logically, if the characteristics of the testing sample do not match those of the normative sample, the resultant scaled score would be by no means valid. IRT remedies these issues by instead considering the probability of getting an item correct (Hambleton, Swaminathan, & Rogers, 1991). Probabilities are easier to compare and manipulate than disparate scores inextricably bound in their own sets of incongruent assumptions. Thus, IRT lends itself readily to the use of test item banks in psychometrics (Boekhall-Timmings, 1990).

As previously emphasized, IRT incorporates the probability that a respondent will correctly answer a specific item. This means that IRT systematically exploits the advantages of having items spread over a broad range of difficulty levels, where more difficult items correspond to higher values of theta. Each item thus has an item information function that corresponds to a specific theta that the respondent is assumed to have if s/he gets that item right (Hambleton, Swaminathan, & Rogers, 1991). If one were to compile the item information functions of all the items in one test to produce a test response function, one would

 ◈ISSN 2094-0734

be able to determine the maximum likelihood estimation of theta. Put simply, this is the most probable theta level corresponding to any specific score.

For every possible score between 0 and the highest possible score, there exist a number of permutations of possible answers, each permutation having its own response functions (Hambleton, Swaminathan, & Rogers, 1991). These are compiled to produce the maximum likelihood estimation of theta for each score. This means that one could simply sum up the number of correct answers in an IRT-approach test and obtain the most likely theta value for that score. This score can stand alone—it needs no further reference to norms and criteria to be of value to the end-user. The maximum likelihood estimation of theta, by definition, also bears the lowest possible standard error of measurement in the function, which considerably simplifies interpretation of data (Hambleton, Swaminathan, & Rogers, 1991).

There are three models typically used in IRT (Anastasi & Urbina, 2005, Hambleton, Swaminathan, & Rogers, 1991). These are the One-Parameter Logistic (1PL, or the Rasch) model, the Two-Parameter Logistic (2PL, or the Birnbaum) model, and the Three-Parameter Logistic Model (3PL) model. The example above simply describes the Rasch model of IRT where the one-parameter measured is item difficulty.

The Birnbaum model is different in that it factors in a second parameter, item discrimination, as well as item difficulty (Anastasi & Urbina, 2005). In this model, one cannot simply sum up correct answers to obtain a theta value—one must also consider which items were correctly answered because the item discrimination value of each item will differentially affect that item's information function, and therefore the maximum likely theta value as well.

The 3PL model foists a third parameter into the fray: Guessing (Anastasi & Urbina, 2005). When a test, like the Graduate Record Exam's Computer Adaptive Test, requires that a response be made before proceeding to the next item, one may have to resort to guessing the correct answer. 3PL models, instead of postulating that those with negative infinity theta (virtually no theta) would have a score of zero, assume that those who don't know will simply guess and likely get 1/k answers correct, where k is the number of possible responses per item.

Which model ought one to use in developing a certain test? Hambleton and others (1991) refer to three assumptions in selecting IRT models. First, the test should be unidimensional (i.e. it should measure only one latent variable). In a factor analysis, one factor should dominate the distribution of eigenvalues in the dataset. This assumption is necessary for all item response theory models. Similarly, there should be invariance between theta and item difficulty.

The second assumption is that of equal discrimination. If the items in the test are unequally discriminating of a person's ability, then it may be necessary to use a 2PL or 3PL model that takes item discrimination into consideration. A heuristic for this would be to see if scores on all the items in a test are equally correlated with the total scores in that test.

The third assumption is the possibility of guessing the correct answer. When guessing is likely to be a factor that will affect scores obtained in a test, the judicious psychometrician would be inclined to use a 3PL model to obtain the appropriate theta values in that test.

 ◈ISSN 2094-0734

The use of IRT has grown substantially over the last couple of decades, finding applications in computer adaptive testing and in high-stakes exams (Hambleton, 2000). IRT applications are typically found in cognitive monotomous (only one answer is positively informative) tests, but recent theoretical developments are showing that polytomous (more than one answer is positively informative) tests can also benefit from IRT approaches to psychometric testing (Bolt, Cohen, & Wollack, 2001). So, how do IRT and CTT comparatively fare in assessment in a single test? Let us refer to an instructive exemplar by Wiberg (2004) where the researcher made use of both CTT and IRT in psychometric assessment.

Wiberg (2004) compared the results of a driving theory license exam in Sweden through CTT and the three models of IRT. In the CTT approach, the test reliability (alpha), the proportion of test-takers who scored correctly on each item (p-value), and the correlation of each test item to the total test score (point-biserial correlation) were used. For IRT, the three aforementioned assumptions were ascertained to be applicable in the dataset, and the 3-parameter logistic model equation was used (2PL and 1PL equations were generalized from the 3PL equation). Furthermore, each item's observed and predicted Item Characteristic Curves (ICC's) were compared through a chi-squared analysis. Finally, each participant's rank on each of the three IRT models was compared as well.

CTT results in the test found a relatively high reliability (alpha = 0.82) and that a scatterplot of p-values and point-biserial correlations showed no apparent relationship between the two variables (i.e. item difficulty and proportion of correct answers in that item did not systematically vary with each other). This implies the likelihood of guessing in the test. For IRT, factor analysis evinced only one factor with a high eigenvalue, supporting the assumption of unidimensionality. Point-biserial correlations varied along a normal distribution. Taken altogether, a 3PL model appeared most appropriate for use in the test.

Continuing with IRT analysis (clearly it is much more complicated than that of CTT), by plotting theta values obtained from easy and difficult items in a scatterplot, one can get an estimate of whether invariance between item difficulty and theta occurred in the test. The scatterplots across IRT models were arguably linear (albeit with quite a bit of noise), which does lend minimal credence to the ability invariance assumption. When item difficulty levels were plotted among low-performing versus high-performing participants (using a cut-off score), the scatterplots for all 3 models, especially 1PL, were linear and demonstrated item difficulty invariance. Guessing estimates were concentrated at a low level for those with high-ability, but varied extensively among those with low ability, indicating that guessing was not invariant across levels of ability.

The last paragraph's results simply show that: (a) The difficulty of an item answered had some effects on theta beyond that expected from the ICC's, hence no invariance in that regard; (b) the items' difficulty level did not vary differently for participants with different ability levels; and (c) low-performing participants varied greatly in their propensity to guess, but high-performing participants did not much vary at all in their propensity to guess.

Proceeding further with the IRT analysis, goodness-of-fit statistics comparing observed and predicted ICC's showed that 53 out of 65 items had to be rejected for the 1PL model to match predicted data. The number of rejected items decreased to 25 for 2PL and 16 in the 3PL model. Furthermore, as expected from an IRT approach, the theta scores obtained from all the respondents fit a normal distribution. Analysis of test information functions (TIF) and their corresponding standard errors found that the 3PL model's TIF was the highest and therefore the most informative. Logically, 3PL's highly-informative TIF had the lowest standard error of measurement. In addition, low-ability and high-ability participants' ranks across IRT models (using scatterplots) were very similar, with middle-performing participants ranked differently only in the 1PL model.

Wiberg (2004) concluded, based on all the aforementioned data, that the 3PL model is optimal for use in the Swedish driving theory license test. Additionally, 3PL item discriminations are similar to CTT point-biserial correlations (hence IRT produces data similar to CTT in addition to its usual benefits). In another analysis, though guessing parameters are likely to vary when there were 4-6 choices in an item, when there were 2 and 3 choices in an item, guessing occurred at a low level of about 0.15 (below chance). This adds valuable information to the measure, in addition to the already plentiful benefits of IRT. CTT would simply assume that the answer is either correct or guessed at a likelihood of 1/k, whereas 3PL IRT incorporates a guessing estimate into the eventual theta. Clearly, CTT and IRT both contribute important information in the assessment of test results.

Item response theory and classical test theory are simply two different approaches to the psychometric science of distilling the thetas, or true values, of the nebulous mental phenomena of interest in psychological science. CTT produces an epistemologically light but weak estimate of theta, whereas IRT, however computationally convoluted, generates an arguably more reliable and informative estimation of theta (rather, the greatest likelihood of theta as represented in a function). Heisenberg's uncertainty principle need not damn the future of psychometrics—IRT has shown that we can indeed cull psychological information with both increasing reliability and increasing information value.

References

Anastasi, A., & Urbina, S. (2005). *Psychological testing (7th Ed.).* NJ: Prentice Hall.

Boekhall-Timmings, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics 15(2),* 129-145.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics 26(4),* 381-409.

Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 98(3),* 357-373.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory.* New York: Sage publications.

Hambleton, R.K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38(9),* 60-65.

Lord, F.M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14(2),* 117-138.

Sijtsma, K., & Junker, B.W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika 33(1),* 75-102.

Wiberg, M. (2004). Classical test theory vs. item response theory: An evaluation of the theory test in the Swedish driving-license test. *Education Measurement 50*. Sweden: Umea University Press. Retrieved September 30, 2008 from http://www.umu.se/edmeas/publikationer/pdf/EM%20no%2050.pdf.