DATA RECOGNITION
DRC
CORPORATION

Accuracy of Test Scores: Why IRT Models Matter

This paper describes different Item Response Theory (IRT) models for both multiple-choice items and constructed-response items. Empirical evidence is presented that indicates that the IRT Three-Parameter Logistic (3PL) and the Two-Parameter Partial Credit (2PPC) models utilized in, for example *TerraNova®, Third Edition*, deliver the most effective representation of student test-taking behavior and produce accurate estimates of student ability. These models are compared with other IRT models, including the Rasch and One- Parameter Partial Credit (1PPC) models.

> *…the IRT Three-Parameter Logistic (3PL) and the Two-Parameter Partial Credit (2PPC) models utilized in, for example* TerraNova, Third Edition, *deliver the most effective representation of student test-taking behavior and produce the most accurate estimates of student ability.*

IRT Models

Item Response Theory uses statistical techniques to model the association between a student's responses to test items and the underlying latent trait (i.e., ability) that is measured by the items. The accuracy of a test score (i.e., the estimation of the underlying ability) depends on how well the IRT model describes this association and fits the test data.

Multiple-Choice Item IRT Models

The most common IRT models for multiple-choice items, where responses of the items are scored dichotomously as right or wrong, are the logistic models with one-, two-, or three- item parameters. The 3PL model depicts the probability of a student with ability θ answering an item *i* correctly as follows:

$$P_i(\theta) = c_i + \frac{1-c_i}{1+e^{-1.7a_i(\theta-b_i)}}$$

This model characterizes a multiple-choice item in terms of three parameters:

- Item difficulty parameter = *b*: Other things being equal, the more difficult (larger *b*) an item, the less likely a student will answer the item correctly.
- Item discrimination parameter = *a*: The item discrimination parameter represents the degree to which responses to the item vary among students with different levels of student ability. Some items have low discriminations, where most students, regardless of their ability level, have about the same probability of getting the item correct. Other items have high discriminations, where there is a strong relationship between students' ability levels and their performance on the item. Items with high discrimination are often desirable because they effectively distinguish among students who differ in ability levels and their performance on an item. Items with high discrimination are often desirable because they effectively distinguish among students who differ in ability.

- Guessing parameter = *c*: The guessing parameter indicates the likelihood of a correct response from a student for whom the item is much too difficult. When a low ability student responds to a very difficult item, the student might guess, and sometimes the student will guess correctly. Because guessing is a part of test taking with multiple choice items, it should be included in the measurement model used to describe test performance.

The graphical display of the modeled association between student ability and the probability of getting an item correct is known as an *Item Characteristic Curve* (ICC). Figure 1 depicts the ICC of a 3PL model item and shows how the change in item parameter values affects the probability curve. Examinee ability (θ) is on the horizontal axis and the probability of getting the item correct is on the vertical axis.

The ICC in the upper left plot shows that the probability of answering an item correctly is a strict increase function of ability. When item difficulty increases (discrimination and guessing unchanged), the curve simply shifts to the right and the probability of a student with a given ability level getting a correct answer on the more difficult item decreases. This can be clearly seen in the upper right plot, where the examinee's chance of getting a correct answer changes from 0.83 with the *b* = -0.5 item to 0.30 with the *b* = 0.1 item.

A more discriminating item better separates two students of different ability. This is depicted in the lower left panel plot, where two students with ability levels of -1.0 and .5 are shown. The difference between the probabilities of these two students answering the more discriminating item correctly (the solid ICC) is A – D. This difference is much larger than the difference between points B and C, the probabilities of the two students answering correctly the less discriminating item (the dashed ICC).

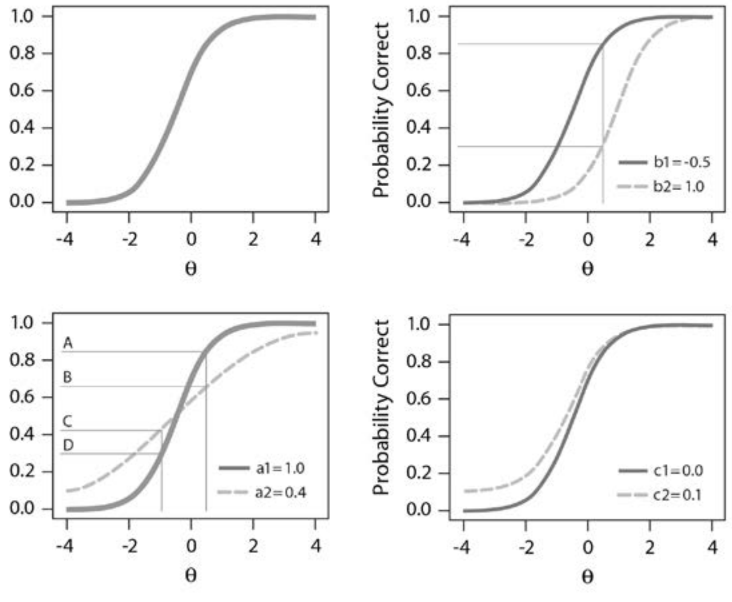

Figure 1: ICCs of 3PL Model

The change of the guessing parameter mostly affects the low ability examinees only, as shown in the lower right plot. An item with a higher guessing parameter allows a higher probability that lower ability examinees will get the item correct.

The 2PL model assumes that there is no guessing involved in students' responses. That is, the guessing parameter *c* = 0, and items differ in discrimination power and difficulty only. The Rasch model, or 1PL model, assumes that there is no guessing and also that all the items are equally discriminating; that is, each has a discrimination parameter *(a) equal to* 1. In the Rasch model, items are assumed to differ only in their difficulty.

Constructed-Response Item IRT Models

Among the commonly used IRT models for constructed-response items are partial credit models with one- or two- item parameters. These models are extensions of the IRT logistic models for multiple-choice items.

The two-parameter partial credit (2PPC) model characterizes an item with item discrimination and item score level difficulty parameters that vary by item score level. The 1PPC model, also known as Masters' partial credit model (Masters, 1982), assumes equal discrimination of all items on a test.

Figure 2 depicts two constructed-response items of three score levels (s = 0, s = 1, and s = 2) modeled by the 2PPC model. Each curve models the relationship between the probability of getting the designated score level on the item and the student's ability level (θ). The s = 0 curve, for example, is the graphical display of probability P(s = 0 | θ). These curves are often called *item category characteristic curve* (ICCC). The two items differ in item discrimination only. The item in the upper panel has a discrimination parameter value of 1.7 and the lower panel item has a smaller discrimination value of 0.75. As the graphs show, the probability curves of the score categories become flatter as the item becomes less discriminating, similar to what happens in the case of multiple-choice items.

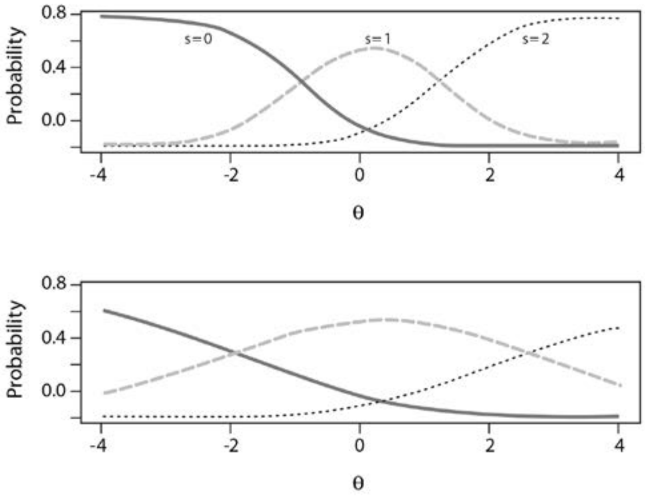

Figure 2: Item Score Level Curve of 2PPC Model

Model Fit and Score Accuracy

Students' test scores are derived from the IRT model used. The accuracy of the scores depends on how closely the IRT model describes the true relationship between students' ability and their item response. Test scores based on a poorly-fitted measurement model.

Test scores based on a poorly-fitted measurement model will be less accurate than those based on a well-fitted measurement model.

To understand which model more closely reflects the true relationship between student's ability and item response, it is important to know what happens in the real world. Guessing is a testing reality. We know that students do indeed guess. Empirical evidence indicates that students guess on multiple-choice items that they either find too difficult or do not have the motivation to consider carefully (Lord, 1980, pp15-17). Empirical data also indicate that low ability students do choose the correct answer to difficult multiple-choice items at a rate that would be expected if they were guessing, and modeling student guessing behaviors has been an important research topic in educational measurement (Woods, 2008; Cao & Stokes, 2008).

Do different items differentiate students differently? Again, empirical evidence indicates clearly that items do have different discrimination powers and yield varying amounts of useful information about student ability (Gleason, 2008; Lord, 1980). For this reason, whenever items differ in their ability to discriminate among students with different abilities, including a discrimination parameter in the IRT model (such as the 3PL or 2PPC IRT models used by DRC) improves the accuracy of the student information provided by the tests.

Note that the 1PL or Rasch model is a special case of the 3PL model. If the 1PL model fits a set of test data, then the 3PL model will automatically fit the data with $c = 0$ and $a = 1$. However, the reverse doesn't hold in many instances; the Rasch model is less likely to fit the data than is the 3PL model. Examinee ability (as reflected by the test scores) will be estimated more accurately if the selected IRT model fits the test data well. Given the reality of examinee guessing, and the importance of accurately modeling an item's discrimination value, the 3PL model should be the first model considered.

This type of scoring, called IRT pattern scoring, yields more accurate estimates of individual student ability than those models based on number-correct scores.

One further point needs to be made when considering the accuracy of the ability estimate or test score. When the Rasch model is used, ability estimates are based only on a student's number-correct score. Thus, all individuals who get the same number- correct score are assigned the same ability estimate, regardless of the characteristics of the particular items that were correctly answered. With the 3PL model, ability estimates can also be based on number-correct scores. The 3PL model, however, allows ability estimates to be based on the student's particular item response pattern. This type of scoring, called IRT pattern scoring, yields more accurate estimates of individual student ability than those models based on number-correct scores. Item-pattern scoring takes into account of items the student correctly answered as well as the characteristics of items missed. It makes sense to give more credit for some questions than others. Extensive analyses of student data have shown that item- pattern scoring produces more accurate scores for individual students than number-correct scoring (Yen, 1984; Yen & Candell, 1991). This additional option of increased accuracy with item-pattern scoring *cannot* occur with the Rasch model.

Final Remarks

One of the most important purposes of an assessment is to obtain valid and accurate estimates of student achievement. The entire assessment process at DRC is designed to attain this goal. From test design and item development through reporting student scores, all efforts are made to ensure that assessments measure what they say they measure, and that the IRT model applied is the model that most accurately reflects real testing behavior. Using the example of *TerraNova, Third Edition*, students are engaged with:

- Real-world problem-solving contexts
- Age- and grade-appropriate language and examples
- Rich graphics and updated item formats
- Authentic literature

Extensive analyses of student data have shown that item-pattern scoring produces more accurate scores for individual students than number-correct scoring

A rigorous process is utilized to develop items that accurately reflect what is taught in today's classrooms. As outlined in the *TerraNova* Technical Reports, every *TerraNova* item went through extensive item content review and field-testing, where the statistical results and characteristics of the items are closely scrutinized. The final *TerraNova* tests are selected to target appropriate ability levels to ensure accurate and fair assessment of student performance.

Different options are available for scoring student responses to obtain estimates of student ability or achievement levels. *TerraNova* employs the most appropriate psychometric models that best reflect students' real testing behavior. The 3PL and 2PPC IRT models accurately define item characteristics and student test taking behavior. The use of these models, along with carefully constructed assessments, results in valid tests that provide accurate test score information that can be used to inform students, parents, and schools about what students know and what they are able to do.

The use of these models, along with carefully constructed assessments, results in valid tests that provide accurate test score information that can be used to inform students, parents, and schools about what students know and what they are able to do.

References

Can, J., & Stokes, S.L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika, v73, n2,* 209-230.

Gleason, J. (2008). An evaluation of mathematics competitions using item response theory. *Notices of The American Mathematical Society,* Vol. 55, No. 1.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, New Jersey.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika,* 47, 149-174.

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

Woods, M.C. (2008). Consequences of ignoring guessing when estimating the latent density in item response theory. *Applies Psychological Measurement, v32, n5,* 371-384.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93-111.

Yen, W. M. & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4,* 209-228.