



ACADEMIC TESTING

ITEM RESPONSE THEORY APPROACH

Despite its popularity, Classical Test Theory (*CTT*) has a number of shortcomings that limit its usefulness as a foundation for modern testing. The emerging role of computing technology in mental testing highlights some of these limitations of *CTT*.

CTT is best suited for traditional testing situations, either in group or individual settings, in which all the members of a target population, e.g., persons seeking college admission, are administered the same or parallel sets of test items. These item sets can be presented to the examinee in either a paper-and-pencil or a computer format. Regardless of format, it is important for the measurement of individual ability that the items in each item set have "difficulties" that match the range of ability or proficiency in the population. In addition, precise estimation of individual ability requires the administration of a "large enough" number of items whose difficulty levels narrowly match the individual's level of ability or proficiency. For heterogeneous populations these requirements of the "fixed length" test result in inefficient and wasteful testing situations that are certainly frustrating to the examinee.

As early as the 1950's, models for mental tests began to appear that addressed these problems with *CTT* and exploited the emergence of computing technology. In fact, a powerful feature of these newer testing models was the ability to choose test items appropriate to the examinee's level of proficiency during the testing session, i.e., to tailor the test to the individual in real time. Today, the

more popular and well developed of these models make up a family of mathematical characterizations of an examinee's test responses known as Item Response Theory (IRT). Although difficult to implement in practice, *IRT* is the formulation of choice for modern testing.

ITEM RESPONSE THEORY

Like CTT, IRT begins with the proposition that an individual's response to a specific test item or question is determined by an unobserved mental attribute of the individual. Each of these underlying attributes, most often referred to as *latent traits* or *abilities*, is assumed to vary continuously along a single dimension usually denoted θ . Under IRT, *both* the test items *and* the individuals responding to them are arrayed on θ from lowest to highest. The position of person i on θ , denoted θ_i , is usually referred to as the person's *ability* or proficiency. The position of item j on θ , usually denoted b_j , is termed the item's *difficulty*. Intuitively, we expect the probability of a correct response to the j -th item to increase monotonically as $\theta_i - b_j$ increases.

In terms of binary scored test items, i.e., items on which responses are designated either correct or incorrect, all IRT models express the probability of a correct response to a test item as a function of θ given one or more parameters of the item. The three most commonly used IRT models are the one-parameter, two-parameter and three-parameter logistic models. Discussion is limited to the three-parameter model since it is the most reasonable model for the common multiple-choice test item and subsumes the one- and two-parameter models.

Three-Parameter Logistic Model

The three-parameter logistic model is given as

$$(1) \quad P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta - b_j)}}$$

where $P_j(\theta)$ is the probability that an examinee with ability θ answers test item j correctly,
 b_j is the item difficulty parameter,
 a_j is the item discrimination parameter, and
 c_j is what we term the "guessing" parameter.

One way to interpret $P_j(\theta)$ is to think of it as the proportion of items that a person of ability θ answers correctly when presented with a very large number of items each having difficulty b_j . Some authors, however, prefer to interpret $P_j(\theta)$ as the proportion of individuals, each with ability θ , that correctly answers item j .

For now, the difficulty of item j , b_j , is defined as the value of θ at which $P_j(\theta) = 0.50$.

The one-parameter logistic model contains only the item difficulty parameter, b_j . The two-parameter model contains both the item difficulty parameter, b_j , and the item discrimination parameter, a_j . While the three-parameter model seems to be the most realistic of the three IRT models in that it acknowledges the chance response through c_j , it is plagued with estimation problems.

Item Characteristic Curve

When plotted, Equation (1) has the appearance of an S-shaped curve having $P_j(\theta) = c_j$ and $P_j(\theta) = 1$ as horizontal asymptotes. For very low values of θ , $P_j(\theta)$ increases slowly. As θ approaches b_j , $P_j(\theta)$ typically increases more rapidly: the rate of increase being a maximum at $\theta = b_j$. As θ becomes

increasingly greater than b_j , the rate at which $P_j(\theta)$ increases slows and approaches 0 for very high values of θ . We now have an alternative definition of item difficulty, b_j , as the point of inflection of the curve relating the probability of a correct item response to ability.

The item discrimination parameter, a_j , is proportional to the slope of the function at $\theta = b_j$. Item discrimination reflects the amount of information about θ contained in the item. (More about item and test information later.) The guessing parameter, $c_j < 1$, is invoked to allow for the possibility of a correct response to item j in the absence of any knowledge or understanding. This is almost always the case with the multiple-choice item format where the correct answer is selected from a small set of plausible answers.

The graph of Equation (1) is called the *Item Characteristic Curve* for item j and denoted ICC_j .

Estimation of Item and Ability Parameters

If the item parameters of the IRT model are known, then the estimation of ability for a sample of examinees is rather straightforward using the method of maximum likelihood. It is common practice to use estimates of the item parameters derived from previous item calibration studies when estimating θ .

The simultaneous estimation of the ability and item parameters of Model (1) is, on the other hand, a much more formidable and problematic task. We describe one procedure for estimating the $N + 3k$ ability and item parameters of the three-parameter logistic model.

Let X_i be a binomial variable taking the value 1 for a correct response to item j and 0 for an incorrect response to item j . Let x_i be the $k \times 1$ vector of binary responses by person i to a set of k items and let $X = (x_1 \dots x_n)$ be the matrix of

such item responses for a random sample of N examinees. The probability of observing \mathbf{X} *prior* to the actual observation of the sample responses is

$$(2) \quad P(\mathbf{X} | \theta, \beta) = \prod_i P(\mathbf{x}_i | \theta_i, \beta) = \prod_i \prod_j E_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}},$$

where $Q_j(\theta) = 1 - E_j(\theta)$, and θ and β are vectors of the fixed unknown parameters of the model. For the three parameter logistic model, the elements of β are the difficulty, discrimination, and guessing parameters of the k items. Expression (2) implies that the responses to any pair of items for fixed θ are independent--a basic assumption of all IRT models.

After the sample observations have been taken, Equation (2) is no longer a probability. Rather, for a given \mathbf{x}_i , Equation (2) can be interpreted as the *likelihood function* $L(\theta, \beta | \mathbf{x}_i)$ of θ and β given \mathbf{x}_i . The value of the likelihood function for a specific set of parameter values is the likelihood that this \mathbf{x}_i would be observed if this set of parameter values was the true set. Obviously, the relative likelihood of \mathbf{x}_i varies as function of the parameters. Those values of the parameters that jointly maximize the likelihood function are termed the *maximum likelihood estimates* of θ and β .

The most common method of estimating the item parameters of the three-parameter logistic model is *Marginal Maximum Likelihood* (MML). MML estimators of the item parameters are obtained by integrating the likelihood function over the ability distribution as

$$(3) \quad L(\beta | \mathbf{x}) = \prod_i \int P(\mathbf{x}_i | \theta, \beta) f(\theta) d\theta,$$

where $f(\theta)$ is the ability density function. $f(\theta)$ may be known a priori or, as in practice, assumed to be the normal density, $N(\mu_\theta, \sigma_\theta^2)$, having $\mu_\theta = 0$ and $\sigma_\theta^2 = 1$. The MML estimates are those values of $\hat{\theta}$ which maximize Equation (3).

Some authors suggest that improved estimates of $\hat{\theta}$ can be obtained by further multiplying $I(\hat{\theta} | X)$ by suitable distributions of the item parameters, $f(a_i)$, $f(b_i)$, and $f(c_i)$. The values that maximize this Bayesian posterior distribution of the parameters are termed the Bayes Modal estimates of $\hat{\theta}$.

Item and Test Information

How precise is the measurement of an individual's ability, θ ? This is the fundamental question that any theory of mental testing must address.

A common statistical index of error in the estimation of a parameter is the variance of the estimator. In the present context, the larger the variance of $\hat{\theta}$ for a given θ , the larger the error associated with $\hat{\theta}$. Conversely, the smaller the variance of $\hat{\theta}$ for a given θ , the greater the precision in estimating ability.

It is customary when using an IRT model to express the variance of $\hat{\theta}$ for a given θ as the reciprocal of the *information function*

$$(4) \quad I(\theta) = \sum_i \frac{[P'_i(\theta)]^2}{P_i(\theta)Q_i(\theta)}$$

where $P'_i(\theta)$ is the first derivative of $P_i(\theta)$ with respect to θ . That is

$$\sigma^2_{(\hat{\theta}|\theta)} = I(\theta)^{-1}$$

Expression (4) is termed the *test information function*. Each of the terms making up the summation in (4) is called the *item information function*. The item and test information functions have some important properties in terms of mental testing.

Turning first to the item information function, we note that this function tends to be symmetric around a maximum attained at $\theta = b_j$ for the one- and two-parameter logistic models and $\theta = b_j + g(a_j, c_j)$ for the three-parameter model. This makes sense in that we expect test items that are either too easy or too difficult for an examinee with ability θ to be less informative than items around $b_j = \theta$. For both the two- and three-parameter models the maximum value of item information is a function of item discrimination, a_j . For all models, including the three-parameter model, items having large values of a_j are more informative than items with smaller a_j values.

Under the three-parameter model, however, the maximum value of item information is a function of both a_j and the guessing parameter, c_j . In this case, as c_j increases the item becomes less informative. Again, this result is reasonable in that we expect items where guessing is a factor to be less informative than items where guessing is absent.

Turning next to the test information function, we first note that the k items contribute independently and, therefore, additively to $I(\theta)$. Along with the fact that $I(\theta)$ is defined for each point on the ability scale, item independence means that items can be chosen--as in adaptive testing--to optimize the estimation of θ .

It is also clear from (4) that test item information depends upon the number of items, k . As k increases, the precision with which ability is estimated increases. In practice, however, we need only administer that number of items needed to attain a predetermined level of precision.

Finally, and in contrast to [Classical Test Theory](#), test information or, conversely, the standard error of measurement for a fixed set of items varies as a function of ability.

Test Construction

The traditional approach to test development based upon [Classical Test Theory](#) centers upon two item statistics: item difficulty (p = the proportion of examinees choosing the correct response) and item discrimination (r_{bi} = the correlation between passing the item and some measure of ability such as the total test score). In general, items are selected whose p values generate the desired test score distribution and that have high item-total test score correlations. The final set of items selected should produce a test score, i.e., ability estimate, having high *reliability*. Test score reliability is used to estimate the error in individual test scores.

The traditional *CTT* approach to test development has several weaknesses. For one, the estimates of item difficulty, item discrimination, and reliability are sample specific. For another, error in examinee test scores is assumed to be constant across all values of ability. In contrast to *CTT*, estimates of IRT item parameters are sample invariant and IRT estimates of error vary as a function of ability. The price of these features of the IRT model is the requirement of very large sample sizes.

The IRT approach to test development centers on the item and test information functions. This seems reasonable in that items should be selected based upon the amount of information each contributes to the amount of information of the test as whole. Once the items have been calibrated, i.e., item parameters have been estimated, item selection is rather straightforward.

Based upon the purpose of the test, e.g., selecting students for a specific academic treatment, specify a *target test information function*. This is the test information function for an optimal set of test items.

Next, select test items with item information functions that meet the requirements of the target function.

Compute the test information function as each item is added to the test.

Continue adding items to the test until the computed test information function approximates, in some acceptable sense, the target test information function.