

An NCME Instructional Module on

# Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development

Ronald K. Hambleton and Russell W. Jones  
*University of Massachusetts at Amherst*

*There are two currently popular statistical frameworks for addressing measurement problems such as test development, test-score equating, and the identification of biased test items: classical test theory and item response theory (IRT). In this module, both theories and models associated with each are described and compared, and the ways in which test development generally proceeds within each framework are discussed. The intent of this module is to provide a nontechnical comparison of classical test theory and item response theory.*

Forty years ago, Dr. Frederic Lord made the important observation that examinee observed scores and true scores are not synonymous with ability scores: Ability scores are more fundamental because they are *test independent* whereas observed scores and true scores are *test dependent* (Lord, 1953). The main idea is that examinees come to a test administration with ability levels or scores in relation to the construct being measured by the test. These ability scores are test independent. However, examinee test scores and corresponding true scores will always depend on the selection of assessment tasks from the domain of assessment tasks over which their ability scores are defined. Examinees will have

lower true scores on difficult tests and higher true scores on easier tests, but their ability scores remain constant over any tests that might be built to measure the construct.

Of course, over time, abilities may change because of instruction and other factors, but at the time of an assessment, each examinee will have an ability score that is defined in relation to the construct, and it remains invariant (i.e., independent) over various samples of assessment tasks that might be used in the assessment. This illustration demonstrates that there is a more fundamental concept than an examinee's true score operating. It is this more fundamental concept that often is of interest in assessment. For example, in computer adaptive testing, in which examinees may see tests of substantially different difficulties, test-dependent scores such as estimated true scores would be of no value in comparing examinees. Ability scores or estimated ability scores that are independent of the particular choice of test items would be of value because they would permit fair norm-referenced comparisons.

Lord (1952, 1953) and a host of psychometricians before him (see for example Gulliksen, 1950) were interested in psychometric theories and models that might lead to descriptions of examinees that would be independent of the particular choice of items or assessment tasks that were used in a test. Also, some psychometricians felt that measurement practices would be enhanced if item and test statistics could be made *sample independent* too. One step in that direction in classical measurement is the preference of biserial correlations over point biserial correlations in estimating item discriminating power because the former are more invariant over examinee samples than the latter (see Lord & Novick, 1968). Basically, however, classical item statistics such as item difficulty (i.e., proportion correct) and item discrimination (i.e., point biserial correlations) and test statistics such as test reliability are dependent on the examinee sample in which they are obtained. Of course, this is not necessarily a problem, and thousands of excellent tests have been constructed in this way (including all important tests up to the end of the 1960s), though special emphasis is placed on obtaining suitable examinee samples for

Ronald K. Hambleton is a Professor of Education and Psychology at the University of Massachusetts, Hills South, Room 152, Amherst, MA 01003. His specializations are item response theory and applications and measurement practices.

Russell W. Jones is a Research Associate at the University of Massachusetts, Hills South, Room 152, Amherst, MA 01003. His specializations are models and applications of item response theory, psychometric methods, and evaluation.

#### Series Information

ITEMS is a series of units designed to facilitate instruction in education measurement. These units are published by the National Council on Measurement in Education. This module may be photocopied without permission if reproduced in its entirety and used for instructional purposes.

obtaining item and test statistics and producing statistically parallel tests.

For measurement specialists who value invariant item and person statistics, one solution lies in the concepts, models, and methods associated with item response theory. This was the point made by Lord in his doctoral thesis and published as a psychometric monograph in 1952 and in an article in 1953 (though at the time the field was known as latent trait theory). The rest is history, as they say. First Allen Birnbaum wrote a series of technical reports in 1957 and 1958 (though his research is more accessible in Birnbaum, 1968), introducing logistic test models and model parameter estimation. Next Georg Rasch (1960) published a book describing several item response models, one of which later became known as the Rasch model (or the one-parameter logistic model), with applications to achievement testing. Later in the 1960s, work by Lord (see Lord, 1968; Lord & Novick, 1968) and by Wright (1968) brought considerable attention to the field of item response theory. Through the 1970s and 1980s the measurement journals (see for example *Applied Psychological Measurement* and the *Journal of Educational Measurement*) were filled with technical advances and applications (see Hambleton, 1989; Hambleton & Swaminathan, 1985; Lord, 1980; Wright & Stone, 1979), and many articles and books for practitioners that described IRT models and their applications were written (see for example Hambleton, 1989; Hambleton, Swaminathan, & Rogers, 1991; Harris, 1989; Loyd, 1988).

Although item response theory and related models and applications have been under study for over 40 years, and extensively studied for the past 25 years, classical test theory and related models have been researched and applied continuously and successfully for well over 60 years, and many testing programs today remain firmly rooted in classical measurement models and methods.

Now in 1993, measurement specialists have a choice of working within a classical test theory framework or an item response theory framework, or even a combination of frameworks. The purposes of this instructional module are (a) to focus attention on the similarities and differences between classical test theory and item response theory and related concepts, models, and methods, and (b) to consider the ways in which tests are developed within the two statistical frameworks. Our main intention is to be descriptive, though we have highlighted advantages and disadvantages of each framework when they have been well-documented in the measurement literature.

### Differences Between Theories and Models

In casual language, the terms “test theories” and “test models” are often used interchangeably. From a technical point of view there is an important difference. Test theories provide a general framework linking observable variables, such as test scores and item scores, to unobservable variables, such as true scores and ability scores. Because of the generality of the specifications, a test theory cannot be shown to be useful or useless. Thus, a test theory that introduces concepts such as true scores, test scores, and error scores cannot be judged as useful or useless until it is fully specified in the form of a particular model. On the other hand, particular test models are formulated within the framework of a test theory and do specify in considerable detail the relationships among a set of test theoretic concepts along with a set of assumptions about the concepts and their relationships. The appropriateness of such models can be evaluated with respect to a particular set of test data. This evaluation may be done by conducting a suitably designed set of empirical investigations or model fit studies. Also, a logical analysis of the model assumptions in relation to the test data can be conducted to address the viability of the model. For example, with a test containing multiple-choice

items and where considerable guessing is expected (perhaps because the test is quite difficult), a test model based on the assumption that true scores and error scores are uncorrelated may not be viable. It would be quite reasonable to expect error scores to be smaller for high-ability examinees and larger for low-ability examinees. Such a finding would be a violation of the assumption; hence, a model incorporating this assumption would, other things being equal, be less useful for the situation described than a model that made a more plausible assumption about the relationship between true scores and error scores.

In the strict sense, models always provide incomplete representations of the test data to which they are fit; hence, with sufficient amounts of test data, they can be falsified, that is, found to be misfitting. Therefore, the meaningful question is not whether a model is correct or incorrect: All models are incorrect in the sense that they provide incomplete representations of the data to which they are applied. The question instead is whether a model fits the data well enough to be useful in guiding the measurement process. Statistical evidence and judgment play important roles in answering the question.

Classical test models are often referred to as “weak models” because the assumptions of these models are fairly easily met by test data. (Though it must be mentioned that not all models within a classical test theoretic framework are “weak.” Models such as the binomial test model, which are based upon a fairly restrictive assumption about the distribution of error scores, are considered “strong models.”) Item response models are referred to as strong models too, because the underlying assumptions are stringent and therefore less likely to be met with test data. For example, the popular one-, two-, and three-parameter logistic models make the strong assumption that the set of items that compose the test are measuring a single common trait or ability. Classical test models do not make such a strong assumption. It is only necessary to assume that the factor structure, whatever it is, is common across parallel forms.

What follows are presentations of classical test theory and item response theory and their related models and concepts, and a comparison of the two statistical frameworks. First, however, a few words about the desirability of test theories and models are offered.

### Importance of Test Theories and Models

Test theories and related models are important to the practice of educational and psychological measurement because they provide a framework for considering issues and addressing technical problems. One of the most important issues is the handling of measurement errors. A good theory or model can help in understanding the role that measurement errors play in (a) estimating examinee ability and how the contributions of error might be minimized (e.g., lengthening a test), (b) correlations between variables (see for example the disattenuation formulas), and (c) reporting true scores or ability scores and associated confidence bands. Different theories and models will handle errors differently. For example, errors might be assumed to be normally distributed in one model, whereas no distributional assumptions about errors are made in another. In one model, the size of measurement errors might be assumed to be constant across the test-score scale (i.e., the standard error of measurement). In another, the size of errors might be assumed to be related to the examinee's true score (i.e., the binomial error model). The specifications about error in a model will have substantial impact on how error scores are estimated and reported.

A good test theory or model can also provide a frame of reference for doing test design work or solving other practical problems. A good test model might specify the precise relationships among test items and ability scores so that careful test

design work can be done to produce desired test score distributions and errors of the size that can be tolerated. For example, in computer adaptive testing, a test model that closely links ability estimates to item statistics is needed to guide the item selection process. Items should be selected at any point in the testing process that provides maximum information about examinee ability. In this application, a model is needed that places persons and items on a common scale (this is done with item response theory models). In this way, at each stage in the computer adaptive testing process, items can be selected that provide the most useful information about examinee ability.

### Classical Test Theory and Related Models

Classical test theory is a theory about test scores that introduces three concepts—test score (often called the observed score), true score, and error score. Within that theoretical framework, models of various forms have been formulated. For example, in what is often referred to as the “classical test model,” a simple linear model is postulated linking the observable test score ( $X$ ) to the sum of two unobservable (or often called *latent*) variables, true score ( $T$ ) and error score ( $E$ ), that is,  $X = T + E$ . Because for each examinee there are two unknowns in the equation, the equation is not solvable unless some simplifying assumptions are made. The assumptions in the classical test model are that (a) true scores and error scores are uncorrelated, (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated. In this formulation, where error scores are defined, true score is the difference between test score and error score. True score is easily shown to be the expected test score across parallel forms. In other formulations of this model (see for example Lord & Novick, 1968), true score is defined as the expected test score over parallel forms, and then the resulting properties of error are derived. In either case, the resulting model is the same and has found wide-spread use in testing practice. Some researchers prefer the latter formulation because it results in defining the concept of central interest, true score, rather than having it obtained as the difference between test score and error score. Gulliksen (1950) actually derived the basic results from the classical test model from both formulations.

About the only definition that is needed is a definition of parallel forms. Parallel forms are defined as tests that measure the same content and for which examinees have the same true score, and where the size of the errors of measurement across forms are equal. Of course, implicit in the definition of parallel forms is the assumption that parallel forms of a test can be constructed. From this modest model, set of assumptions, and definition, most of the test formulas and results that are sprinkled through tests and measurement texts were generated. To this day, many important tests are constructed from the classical test model. Important results that derive from the model (such as the generalized Spearman-Brown formula, the formula for linking test length to test validity, and the disattenuation formulas) are well-known and commonly used in testing practice.

To derive other models within the framework of classical test theory, researchers have moved in many directions including dropping or revising one or more of the basic assumptions, or adding distributional assumptions about error and true scores. For example, the assumption that the distribution of errors follows a binomial distribution (i.e., the binomial test model) or a normal distribution has been common (see Crocker & Algina, 1986). The binomial test model has been particularly useful in addressing technical problems that have arisen with criterion-referenced measurement. This model is used in determining test length, assessing reliability, and assessing mastery. With still other models, the definition of parallel forms has been weakened (i.e., the requirement that true scores should be

equal across parallel forms has been replaced by a model in which all that is required is that true scores across parallel-forms be linearly related). Still other researchers have extended classical test models to specify in considerable detail the error score by identifying components of error (such as errors resulting from the scorer, item format, and the particular test administration) and then designing studies to assess these components and their impact on test score variance and test reliability (see for example Shavelson & Webb, 1991). In sum, the field of classical test theory is served by a variety of models.

Most of the work in classical test theory has focused on models at the test-score level (in contrast to item response theory). That is, the models have linked test scores to true scores rather than item scores to true scores. However, item statistics that represent item difficulty (often denoted  $p$ ) and item discriminating power (often denoted  $r$ ) have been developed, and their connection to test statistics such as test-score mean, standard deviation, and reliability are well-known and are used in the test development process to produce tests with the desired statistical properties. Readers are referred to excellent chapters by Henrysson (1971) and Millman and Greene (1989) for specifics on how item statistics are defined and used in test development. Suffice it to say here that these item statistics (and their variations) have proven valuable in the test development process. Still, one main shortcoming is that they are sample dependent, and this dependency reduces their utility. They are most useful when the examinee sample is similar to the examinee population for whom the test is being developed. To the extent that the sample differs in some unknown way from the population, and this could easily happen in a field test, the utility of the item statistics may be reduced. The use of some “anchor items” in a field test that also appeared in an actual test administration can be used to partially resolve sampling problems but relationships are typically nonlinear, which complicates any analyses.

Advantages of many classical test models are that they are based on relatively weak assumptions (i.e., they are easy to meet in real test data) and they are well-known and have a long track record. On the other hand, both person parameters (i.e., true scores) and item parameters (i.e., item difficulty and item discrimination) are dependent on the test and the examinee sample, respectively, and these dependencies can limit the utility of the person and item statistics in practical test development work and complicate any analyses.

### Item Response Theory and Related Models

Item response theory is a general statistical theory about examinee item and test performance and how performance relates to the abilities that are measured by the items in the test. Item responses can be discrete or continuous and can be dichotomously or polychotomously scored; item score categories can be ordered or unordered; there can be one ability or many abilities underlying test performance; and there are many ways (i.e., models) in which the relationship between item responses and the underlying ability or abilities can be specified. Within the general IRT framework, many models have been formulated and applied to real test data (see Hambleton, 1989, for a review). In this module, only a few of the models that (a) assume a single ability underlies test performance, (b) can be applied to dichotomously scored data, and (c) assume the relationship between item performance and ability is given by a one-, two-, or three-parameter logistic function will be considered. Typically, two assumptions are made in specifying IRT models: One relates to the dimensional structure of the test data, and the other relates to the mathematical form of the item characteristic function or curve (denoted ICC).

Figure 1 shows the general form of item characteristic functions with the three-parameter logistic model. Item charac-

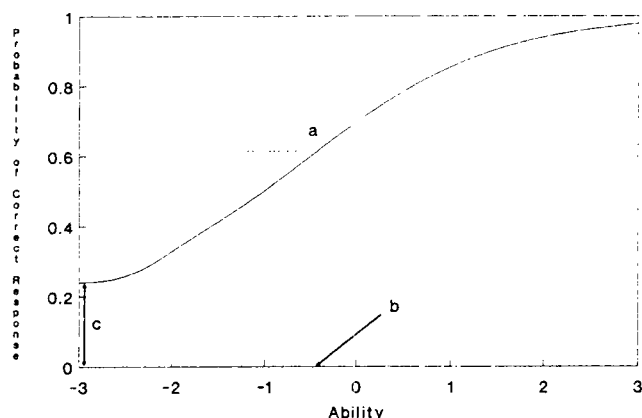


FIGURE 1. A three-parameter logistic model item characteristic curve

teristic functions are generated from the expression

$$P_i(\theta) = c_i + (1 - c_i)[1 + e^{-Da_i(\theta - b_i)}]^{-1}, \quad i = 1, 2, \dots, n,$$

which serves as the mathematical model linking the observable data (item performance) to the unobservable data (ability).  $P_i(\theta)$  gives the probability of a correct response to item  $i$  as a function of ability (denoted  $\theta$ ). The symbol “ $n$ ” is the number of items in the test. The  $c$  parameter in the model is the height of the lower asymptote of the ICC and is introduced into the model to account for the performance of low-ability examinees on multiple-choice test items. This parameter is not needed in the model with free-response data. The  $b$  parameter is the point on the ability scale where an examinee has a  $(1 + c)/2$  probability of a correct answer. The  $a$  parameter is proportional to the slope of the ICC at the point  $b$  on the ability scale. In general, the steeper the slope, the higher the  $a$  parameter. The item parameters,  $b$ ,  $a$ , and  $c$ , are correspondingly referred to as the item difficulty, item discrimination, and pseudoguessing parameters. The “ $D$ ” in the model is simply a scaling factor. By varying the item parameters, many S-shaped curves or ICCs can be generated to fit actual test data. Simpler logistic test models can be obtained by setting  $c_i = 0$  (the two-parameter model) or setting  $c_i = 0$  and  $a_i = 1$  (the one-parameter model). Thus, three different logistic models may be fit to the test data. A typical set of ICCs is shown in Figure 2. The corresponding item statistics for the items appear in Table 1. For more details, readers are referred to Hambleton (1989) or Harris (1989).

Some of the flexibility of item response theory arises because the models link item responses to ability, and item statistics are reported on the same scale as ability. This is not the case in classical test theory. The flexibility comes through knowing precisely where an item is doing its best measurement on the ability scale, and knowing the exact relationship between item performance and ability.

One useful feature is that of the *test characteristic function* represented in Figure 3. It is the sum of the item characteristic functions that makes up a test and can be used to predict the scores of examinees at given ability levels. If the test is made up of test items that are relatively difficult, then the test characteristic function is shifted to the right and examinees tend to have lower expected scores on the test than if easier test items are included. Thus, it is possible through the test characteristic function to explain how it is that examinees with a fixed ability can perform differently on two tests measuring the same ability, apart from the ubiquitous error scores. The test characteristic function connects ability scores in item response theory to true scores in classical test theory because

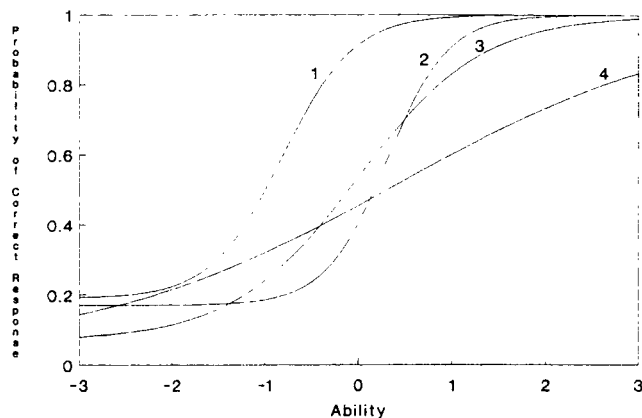


FIGURE 2. Three-parameter ICCs for four test items

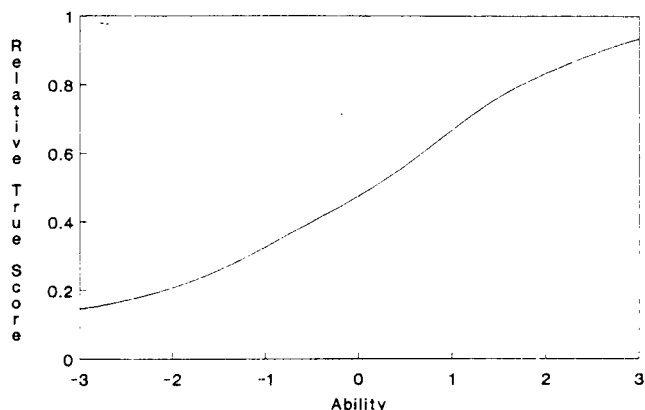


FIGURE 3. Relationship between ability and true score

an examinee’s expected test score at a given ability level is by definition the examinee’s true score on that set of test items.

Another feature of item response theory models is the existence of item information functions. In the case of the simple logistic models, item information functions show the contribution of particular items to the assessment of ability. In general, items with high discriminating power contribute more to measurement precision than items with lower discriminat-

**Table 1**  
*Item Statistics for the 12-Item Bank*

Item	Item statistic		
	$b$	$a$	$c$
1	-0.81	1.55	.19
2	0.31	1.78	.17
3	0.00	0.88	.07
4	0.45	0.36	.04
5	-2.41	0.46	.08
6	-1.30	1.24	.05
7	2.23	0.61	.03
8	1.00	1.85	.08
9	1.40	1.01	.03
10	-1.95	0.85	.14
11	2.56	1.44	.13
12	1.10	1.33	.08

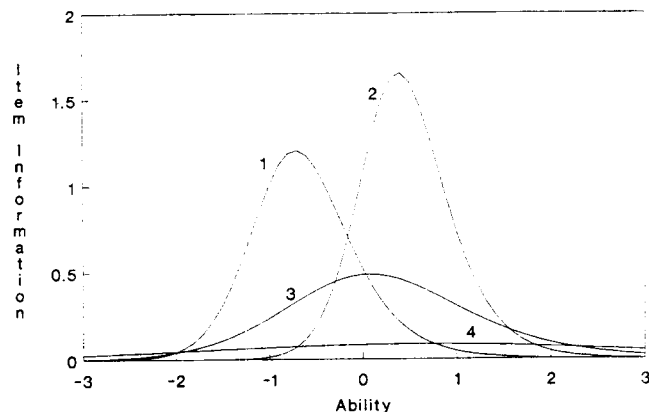


FIGURE 4. Item information functions for four test items

ing power, and items tend to make their best contribution to measurement precision around their  $b$  value on the ability scale. Figure 4 shows the item information functions that correspond to the items shown in Figure 2. Notice for example that Item 2 is more difficult than Item 1; hence, the item information function for Item 1 is centered at a lower ability level than the item information function for Item 2. Also, because Items 3 and 4 are less discriminating than Items 1 and 2 (see Figure 2), the corresponding item information functions are lower than for Items 1 and 2.

Another special feature of item response models is the concept of a *test information function*, denoted  $I(\theta)$ . It is the sum of item information functions in a test and provides estimates of the errors associated with (maximum likelihood) ability estimation, specifically,

$$SE(\theta) = \frac{1}{[I(\theta)]^{1/2}}$$

This means that the more information provided by a test at a particular ability level, the smaller the errors associated with ability estimation. Figure 5 provides the test information function for the 12 items shown in Table 1. This test provides more precise estimates of ability near  $\theta = 1.0$  than other points on the ability scale. The presence of item and test information functions substantially alters the ways in which tests are constructed within an item response theory framework. These points are developed later in the module.

Item and test characteristic functions and item and test information functions are integral features of item response

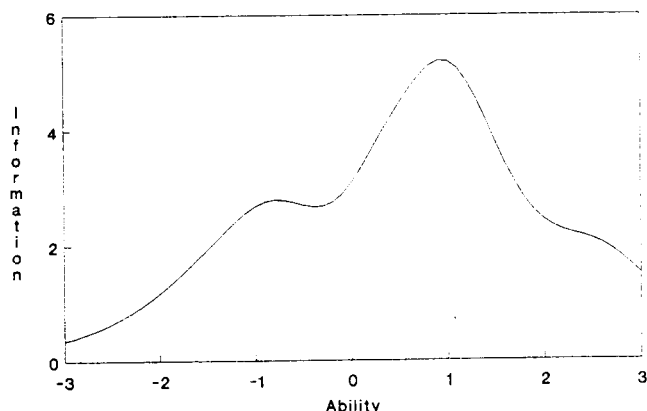


FIGURE 5. Information function for the 12-item test

theory models, and they are immensely useful. However, the essential property of these functions is what is important; that is, model-parameter invariance. Figure 6 represents the situation highlighting item-parameter invariance. Notice that, for Groups 1 and 2, the same item characteristic function applies. For each ability level, there is a probability of a correct response. Of course, that probability does not and should not depend on the number of examinees in each group at that ability level. In that sense, the ICC applies equally well to both groups. Were classical statistics for the item shown in Figure 6 computed, the item would be easier and more discriminating in Group 2 than in Group 1. It can also be shown that person parameters or abilities are estimated independently of the particular test items, and this is accomplished by incorporating the item statistics into the ability estimation process. Of course, the property of model parameter invariance is only obtained with models that fit the test data to which they are applied.

In recent years, the number of item response theory models has expanded to handle the diverse types of data that are being produced by more performance-oriented tests. Models to handle polychotomous data (i.e., the partial credit model and the graded-response model) as well as models to handle multidimensional data are now being widely researched, and the former models in particular are being used with performance data, such as writing assessments.

Like the classical test models, item response theory models are in wide use in test development, equating test scores, identifying item bias, and scaling and reporting scores. These models are being used by many national and state organizations, and even large school districts.

### Comparison of Theories and Models

In the two previous sections many points of similarity and difference between classical test theory and item response theory were highlighted. Table 2 provides a summary organized around eight areas. A few additional points follow.

The relationship between item difficulty and discrimination parameters in the classical test model and the two-parameter

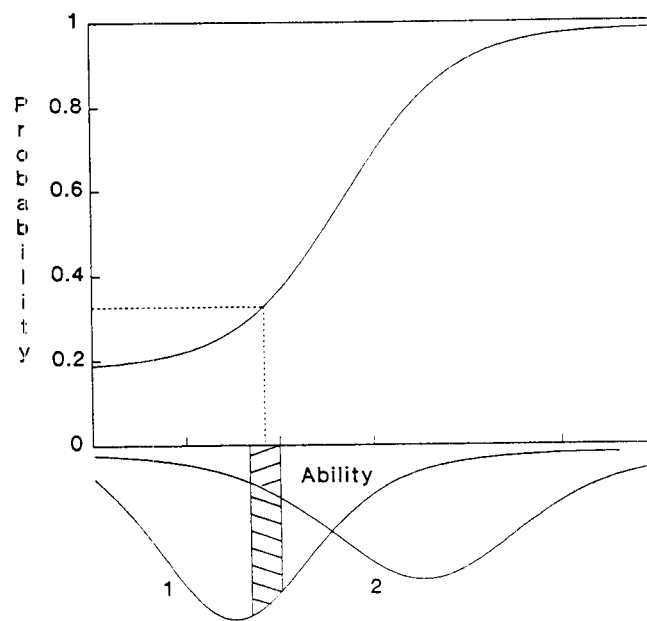


FIGURE 6. Item characteristic curve and ability distributions for two groups of examinees

**Table 2****Main Differences Between Classical and Item Response Theories and Models**

Area	Classical test theory	Item response theory
Model	Linear	Nonlinear
Level	Test	Item
Assumptions	Weak (i.e., easy to meet with test data)	Strong (i.e., more difficult to meet with test data)
Item-ability relationship	Not specified	Item characteristic functions
Ability	Test scores or estimated true scores are reported on the test-score scale (or a transformed test-score scale)	Ability scores are reported on the scale $-\infty$ to $+\infty$ (or a transformed scale)
Invariance of item and person statistics	No—item and person parameters are sample dependent	Yes—item and person parameters are sample independent, if model fits the test data
Item statistics	$p, r$	$b, a$ , and $c$ (for the three-parameter model) plus corresponding item information functions
Sample size (for item parameter estimation)	200 to 500 (in general)	Depends on the IRT model but larger samples, i.e., over 500, in general, are needed

logistic model is discussed by Lord (1980). He shows that, under certain conditions (such as examinee performance not being affected by guessing), the item-test biserial correlation used within the framework of classical measurement theory and the item discrimination parameter of item response theory are approximately monotonically increasing functions of each other. This relationship may be represented as:

$$a_i \cong \frac{r_i}{\sqrt{1 - r_i^2}}$$

where

$a_i$  = item discrimination parameter value for item  $i$  used in IRT and

$r_i$  = item biserial correlation.

The relationship is approximate rather than accurate as a consequence of the different distributions and assigned scores of the two models. The number correct score ( $X$ ) of classical test theory and the ability score ( $\theta$ ) of item response theory have distributions with different shapes, and the relationship between  $X$  and  $\theta$  is nonlinear. Furthermore, the total test score  $X$  is subject to errors of measurement, whereas the ability score  $\theta$  is not (Lord, 1980).

Lord describes a similar monotonic relationship between  $p_i$  and  $b_i$  when all items are equally discriminating (such as in the Rasch model) so that as  $p_i$  increases  $b_i$  decreases. If items have unequal discrimination values, then the relationship between  $p_i$  and  $b_i$  will depend on  $r_i$  (Lord, 1980). This relationship may be represented as:

$$b_i \cong \frac{\gamma_i}{r_i}$$

where

$b_i$  = item difficulty parameter value for item  $i$  used in IRT and

$\gamma_i$  = normal deviate corresponding to the ability score beyond which  $p_i$  of the examinee sample fall.

Perhaps the most important distinction between classical

and modern test theories is that inherent within item response theory is the property of invariance of both item parameters and ability parameters. The consequences of this property are (a) those parameters that characterize an examinee are independent of the test items from which they are calibrated and (b) those parameters that characterize an item are independent of the ability distribution of the set of examinees (Hambleton, Swaminathan, & Rogers, 1991).

Classical test theory has a number of important limitations. First and foremost is that the two statistics (item difficulty and item discrimination) that form the cornerstones of many classical test theory analyses are group dependent. Thus, the  $p$  and  $r$  values, so essential in the application of classical test models, are entirely dependent on the examinee sample from which they are obtained. In terms of discrimination indices, this means that higher values will tend to be obtained from heterogeneous examinee samples, and lower values from homogeneous examinee samples. In terms of difficulty indices, higher values will be obtained from examinee samples of above-average ability and lower values from examinee samples of below-average ability (Hambleton, 1989).

Another limitation of classical test theory is that scores obtained by classical test theory applications are entirely test dependent. Consequently, test difficulty directly affects the resultant test scores. This is an important shortcoming because the practical constraints of measurement practice frequently necessitate that examinees from a single population be compared using results obtained from different test items as a result of having been administered different forms of the same test, or at least different subtests. Indeed, classical test theory may be described as "test based," whereas IRT may be described as "item based." The true-score model upon which much of classical test theory is based permits no consideration of examinee responses to any specific item. Consequently, no basis exists to predict how an examinee, or a group of examinees, may perform on a particular test item. Conversely, item response theory allows the measurement specialist greater flexibility. A broader range of interpretations may be made at the item level. Thus, item response theory permits the measurement specialist to determine the probability of a particular examinee correctly answering any given item. This

has obvious advantages if a test developer needs to know the characteristics of the test scores of one or more examinee populations. Similarly, if it is necessary to design a test with particular inherent characteristics for a specific examinee population, item response models permit the test developer to do just that (Hambleton, Swaminathan, & Rogers, 1991). The need to build such tests is common: for example, a test built to discriminate among less-able students to select candidates for limited special-needs resources or a test built to discriminate among more-able students for the award of a scholarship. In particular, this property of item response theory is invaluable for certain modern testing applications, such as computerized adaptive testing.

Item response models have technical and practical shortcomings, too. On the technical side, item response theory models tend to be complex, and model parameter estimation problems tend to arise in practice. Model fit too can be a problem—it is still not completely clear how problems of model fit should be addressed, especially problems that relate to test dimensionality. On the practical side, almost regardless of application, the technical demands tend to be more complex than the demands that arise with classical models. The one-parameter item response theory model certainly is more straightforward to apply than the other item response theory models (and the software, in general, is user-friendly). On the other hand, questions arise about the fit of the one-parameter model because of the restrictiveness of the model assumptions.

An awareness of the shortcomings of classical test theory and the potential benefits offered by item response theory has led some measurement practitioners to opt to work within an item response theory framework. The reason for this change of emphasis by the psychometric and measurement community from classical to item response models is as a consequence of the benefits obtained through the application of item response models to measurement problems. These benefits include:

1. Item statistics that are independent of the groups from which they were estimated.
2. Scores describing examinee proficiency that are not dependent on test difficulty.
3. Test models that provide a basis for matching test items to ability levels.
4. Test models that do not require strict parallel tests for assessing reliability.

Benefits obtainable through the application of classical test models to measurement problems include:

1. Smaller sample sizes required for analyses (a particularly valuable advantage for field testing).
2. Simpler mathematical analyses compared to item response theory.
3. Model parameter estimation is conceptually straightforward.
4. Analyses do not require strict goodness-of-fit studies to ensure a good fit of model to the test data.

## Test Development

Figure 7 displays the typical steps in test development. Important differences between test development using classical and item response measurement theories occur at Steps 3, 5, and 9. Discussion within this module will focus on item analysis and the selection of items for a test (see also Yen, 1983).

### Classical Test Theory

*Item analysis.* When employing the standard test development techniques of classical test theory, item analysis consists of (a) determining sample-specific item parameters by employing simple mathematical techniques and moderate sample sizes, and (b) deleting items based on statistical criteria. Standard item analysis techniques involve an assessment of item difficulty and discrimination indices and item distractors.

1. Preparation of test specifications
2. Preparation of the test item pool
- \*3. Field testing the items
4. Revision of the test items
- \*5. Test development
6. Pilot testing
7. Final test development
8. Test administration (for norming and technical data)
- \*9. Technical analyses (e.g., compiling norms, setting standards, equating scores, reliability and validity studies)
10. Preparation of administrative instructions and technical manual
11. Printing and distribution of tests and manuals.

\*Important differences in test development using classical test theory and item response theory occur at Steps 3, 5, and 9.

FIGURE 7. Steps in test development

Because item statistics depend to a great extent on the characteristics of the examinee sample used in the analysis, an important concern of test developers applying classical test theory is that the examinee sample should be representative of the overall population for whom the test is intended. Heterogeneous samples will, generally, result in higher estimates of item discrimination indices as measured by point-biserial or biserial correlation coefficients, whereas item difficulty estimates rise and fall with high- and low-ability groups, respectively. Despite the inherent difficulty of obtaining a representative sample, an advantage of this approach to item analysis is that item statistics can be accurately calibrated on examinee samples of modest size.

Detection of poor items (at least for norm-referenced tests) is quite straightforward and is basically accomplished through careful study of item statistics. A poor item is identified by an item difficulty value that is too high or too low, or a low item-total score correlation. It is appropriate to point out that classical item analysis procedures, together with an analysis of distractors, have the potential to provide the test developer with invaluable information concerning test item quality regardless of which measurement model is applied in the later stages of test development.

*Item selection.* When applying test development techniques, in addition to concerns for content validity, items are selected on the basis of two characteristics: item difficulty and item discrimination. An attempt is always made to choose items with the highest discrimination parameters. The choice level of difficulty is usually governed by the purpose of the test and the anticipated ability distribution of the group for whom the test is intended. For example, it may be the case that the purpose of a test is to select a small group of high-ability examinees for the award of a scholarship. In this situation, items are generally selected that are quite difficult for the population at large. Most norm-referenced achievement tests are commonly designed to differentiate examinees with regard to their competence in the measured areas. That is, the test is designed to yield a broad range of scores maximizing discriminations among all examinees taking the test. When a test is designed for this purpose, items are generally chosen to have a medium level and narrow range of difficulty.



**Item analysis.** When employing item response theory, item analysis consists of (a) determining sample-invariant item parameters using relatively complex mathematical techniques and large sample sizes, and (b) utilizing goodness-of-fit criteria to detect items that do not fit the specified response model. The property of sample invariance inherent within IRT means that test developers do not need a representative sample of the examinee population to calibrate test items. They do, however, need a heterogeneous and large examinee sample to insure proper item parameter estimation. As can be seen from Figure 6, even when examinee samples differ, the test developer is able to use the principles of IRT to estimate the same ICC regardless of the examinee sample used in the item calibration process. However, the test developer using IRT is faced with a different problem. Because IRT requires larger sample sizes to obtain good item parameter estimates, the test developer must ensure that the examinee sample is of sufficient size to guarantee accurate item calibration.

The detection of poor items using item response theory is not as straightforward as when classical test theory is used. Items are generally evaluated in terms of their goodness-of-fit to a model using a statistical test or an analysis of residuals. Although this is not the appropriate article to discuss goodness-of-fit procedures, it is important to emphasize that an adequate fit of model-to-data is essential for successful item analysis; otherwise, items may appear poor as an artifact of poor model fit. Readers wishing to learn more about goodness-of-fit techniques are referred to Hambleton, Swaminathan, and Rogers (1991, chapter 4) and Hambleton and Murray (1983). Poor items are usually identified through a consideration of their discrimination indices (the value of  $a_i$  will be a low positive or even negative) and difficulty indices (items should be neither too easy nor too difficult for the group of examinees to be assessed).

**Item selection.** As is the case with classical test theory, item response theory also bases item selection on the intended purpose of the test. However, the final selection of items will depend on the information they contribute to the overall information supplied by the test. A particularly useful feature of the item information functions used in IRT test development is that they permit the test developer to determine the contribution of each test item to the test information function independently of other items in the test. Lord (1977) outlined a procedure, originally conceptualized by Birnbaum (1968), for the use of item information functions in the test building process. Basically, this procedure entails that a test developer take the following four steps:

1. Describe the shape of the desired test information function over the desired range of abilities. Lord (1977) calls this the *target information function*.
2. Select items with item information functions that will fill up the hard-to-fill areas under the target information function.
3. After each item is added to the test, calculate the test information function for the selected test items.
4. Continue selecting test items until the test information function approximates the target information function to a satisfactory degree.

For example, suppose a test developer wished to build a test to fill the target information function shown in Figure 5. The first step would be to specify this information function as the target information function. The next step would be to add an item to the test, filling up the hard-to-fill areas first. Item 2 in Figure 4 would be a good choice, followed by Item 1. The test information function should be recalculated after the addition of each item so as to identify the specific information required to complete the test. By following this procedure, and selecting the 12 items in Table 1, the test developer would create the

ideal test to match the target information function shown in Figure 5. Content validation considerations are monitored during the item selection process.

This procedure allows the test developer to build a test that will precisely fulfill any set of desired test specifications. Thus, it is possible to build a test that "discriminates" well at any particular region on the ability continuum. That is to say, if we have a good idea of the ability of a group of examinees, test items can be selected so as to maximize test information in the region of ability spanned by the examinees being tested. Of course, this optimum selection of test items will contribute substantially to the precision with which ability scores are estimated. Furthermore, with criterion-referenced tests, it is common to observe lower test performance on a pretest than on a posttest. Given this knowledge, a test instructor should select easier items for the pretest and more difficult items for the posttest. Then, for both testing administrations, measurement precision will have been maximized in the ability region where the examinees would most likely be located. Moreover, because items on both tests measure the same ability, and ability estimates are independent of the particular choice of test items, the instructor can measure growth by subtracting the pretest ability estimate from the posttest ability estimate.

### Self-Test

1. a. What are the item statistics used in the classical test theory model and the two-parameter logistic model?  
b. How are they related?
2. a. List the expected benefits when an item response model is applied to a measurement problem.  
b. List the expected benefits when a classical test model is applied to a measurement problem.
3. a. What are the basic assumptions of the classical test theory model?  
b. What are the basic assumptions of item response theory models?
4. Estimate the item difficulty and pseudoguessing parameter values for the four items in Figure 2.
5. a. Suppose Figure 4 presents the item information functions for four reading items. If you were a test developer tasked with selecting an item for use in a test for a remedial reading population, which of the four items would you choose for inclusion in the test? Why?  
b. Which of the four items would you choose if you were building a test to discriminate among more able readers? Why?
6. a. An item has  $b = 1.0$ ,  $a = 0.8$ , and  $c = 0.2$ . Calculate  $P(\theta)$  at  $\theta = -3, -2, -1, 0, 1, 2$ , and  $3$ .  
b. Repeat this procedure for a second item:  $b = 0.5$ ,  $a = 0.6$ ,  $c = 0.15$ .  
c. Sketch the ICCs for the two items (that is, plot the probability of success as a function of ability for each test item). Which of the two items would an examinee with an ability of 1.0 have the greatest probability of answering correctly?
7. a. A subtest is composed of the five items whose item information values at seven ability levels are shown as follows:

Item	Ability level						
	-3	-2	-1	0	1	2	3
1	.01	.25	.50	.70	.83	.65	.40
2	.05	.35	.70	.85	.70	.35	.05
3	.10	.30	.40	.45	.10	.05	.01
4	.10	.30	.40	.45	.40	.30	.10
5	.40	.65	.83	.70	.50	.25	.01



- What is the subtest information function at each ability level?
- b. If practical test construction constraints required the size of the subtest to be reduced to only four items, which item should be removed?

### Answers to Self-Test

- Classical test theory model: item difficulty ( $p$ ) and item discrimination ( $r$ ).  
Two-parameter logistic model: item difficulty ( $b$ ) and item discrimination ( $a$ ).
  - When  $r$  is high, so is  $a$ ; when  $r$  is low, so is  $a$ ; there is a monotonically increasing relationship between the two. On the other hand,  $p$  and  $b$  are inversely related. If ability is normally distributed, and the two-parameter logistic model fits the test data, there are specific formulas linking  $r$  to  $a$  and  $p$  to  $b$ . See Lord and Novick (1968, p. 378) for the formulas.
- (a) Item statistics are independent of the groups from which they were estimated; (b) scores describing examinee proficiency are not dependent on test difficulty; (c) test models that provide a basis for matching test items to ability levels where they function best.
  - (a) Well-known models that have been useful for a long time; (b) weak assumptions, hence the useful formulas are widely applicable.
- In what is often called the *classical test model*, basically, there are three assumptions, in addition to the model itself:

$$X = T + E$$

and

$$\rho_{TE} = 0, \quad \bar{E} = 0, \quad \text{and} \quad \rho(E_1, E_2) = 0$$

- Assumptions are made about (a) dimensionality (usually, that the test measures one dimension or trait), and (b) the mathematical form of the item characteristic function (or curve). (For dichotomously scored items, logistic functions are common.)
- The item statistics for the four items in Figure 2 are:

Item	Item difficulty ( $b$ )	Pseudoguessing parameter ( $c$ )
1	-0.81	.19
2	0.31	.17
3	0.00	.07
4	0.45	.14

- Item 1. Of the four items, Item 1 provides the most information at lower ability levels.
  - Item 2. Of the four items, Item 2 provides the most information at higher ability levels.
- Here are the probabilities:

Item	Ability level						
	-3	-2	-1	0	1	2	3
6a	.203	.213	.249	.363	.600	.837	.951
6b	.173	.212	.301	.469	.681	.849	.938

- Item 6b. See Figure 8.
  - The subtest information function at the seven ability levels is:

	Ability level						
	-3	-2	-1	0	1	2	3
Information	.66	1.85	2.83	3.15	2.53	1.60	.57

- Item 3. This item provides the least amount of information over the ability range of interest.

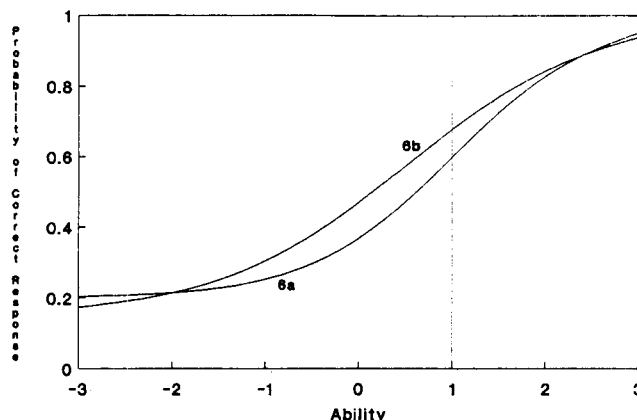


FIGURE 8. Plot of item characteristic curves for questions 6a and 6b

### References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: Macmillan.
- Hambleton, R. K., & Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.
- Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, 8, 35-41.
- Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph*, 7.
- Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, 517-548.
- Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, 28, 989-1020.
- Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Loyd, B. H. (1988). Implications of item response theory for the measurement practitioner. *Applied Measurement in Education*, 1(2), 135-143.
- Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: Macmillan.

- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems* (pp. 85–101). Princeton, NJ: Educational Testing Service.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Yen, M. W. (1983). Use of the three-parameter logistic model in the development of a standardized achievement test. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 123–141). Vancouver, British Columbia, Canada: Educational Research Institute of British Columbia.