

ITEM RESPONSE THEORY

Item response theory (IRT) is a family of statistical procedures for analyzing and describing test performance. It has three major characteristics that distinguish it from classical test theory: (a) It focuses on performance on individual items, rather than only on intact tests; (b) it describes item performance at each level of student ability; and (c) it is model based. Researchers utilize IRT to analyze students' item performance data from one testing situation, describe it succinctly, and make predictions about item and test performance in other situations. The use of IRT takes advantage of generalizable information in a flexible manner to increase the efficiency and usefulness of the measurement process.

During the past 20 years, the use of IRT in education has increased tremendously. Almost every major publisher of educational tests uses IRT in some way, as do a substantial number of local educational agencies. Research on the properties and characteristics of IRT also has increased dramatically. Much of this research has focused on detailing the circumstances under which the theoretical advantages of IRT are fulfilled in practice.

Conceptual Foundation

The idea of generalizing or predicting from one testing occasion to others is one of long standing in educational measurement. For example, the proportion of stu-

dents passing an item (*p-value* or p_i) is a traditional description of item difficulty. Using *p-values* obtained from one group of students allows items to be rank ordered from easiest to most difficult and the prediction made that they will have the same rank ordering for other groups of students. Although *p-values* are very useful for predicting rank orders of item difficulties, they do not include a way of predicting how much the numerical values of the *p-values* will change when more or less able groups of students are tested.

Going beyond traditional procedures, IRT describes item difficulty in a manner that is stable over groups of students and states how this stable function interacts with the ability level of the group. This process permits detailed predictions of how much *p-values* will change when different students are tested. Going further than just describing difficulty, IRT can describe other statistical properties as well, permitting detailed predictions of many items and test characteristics.

In order for IRT to produce these descriptions and predictions efficiently, a statistical model is required. A model makes simplifying assumptions about the important factors influencing item performance. The core of each IRT model is a formula that defines the probability of a student's correct response to an item as a function of the ability of the student and properties of the item. This function is called the *item characteristic function*, and a graph of it is called the *item characteristic curve* (ICC). A simplified example is useful for describing its conceptual basis.

Imagine that a 20-item test is given to a group of students. The proportion of students with each number-correct score that passes Item 1 could be obtained and plotted as in Figure 1(a); performance on Item 2 also appears there. This information is *conditional*, which means that it presents the chance of passing each item conditional on—or given that—a student has attained each total test score. From Figure 1(a), a student's total

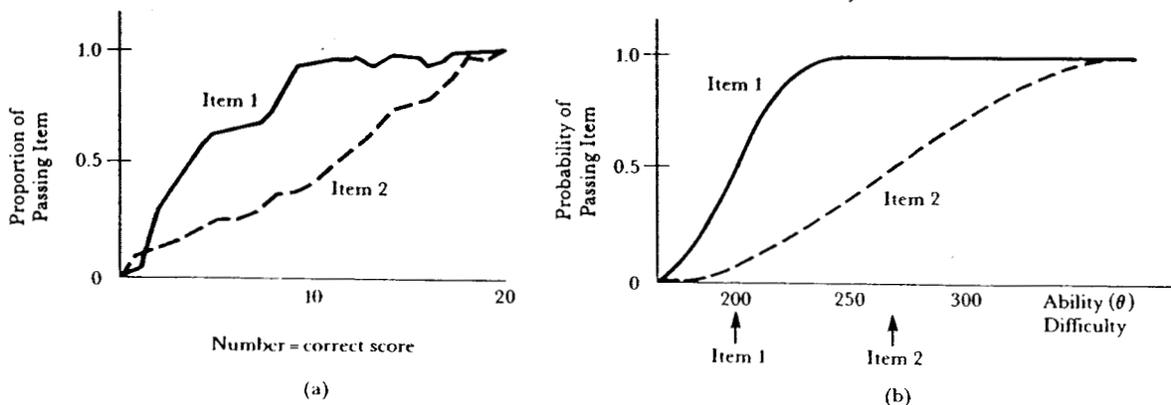
test score can be used to predict his or her item performance. For example, a student with a test score of 10 would be predicted to have a 95% chance of passing Item 1 and a 50% chance of passing Item 2.

Although Figure 1(a) provides a great deal of information, it has disadvantages. Maintaining all the possible data points for each item in order to describe that item would be cumbersome, and the minor irregularities in the curves are likely to be unimportant. Also, in Figure 1(a), item performance is described conditional on a number-correct score that includes that item; an item's conditioning function will change, depending on the number and characteristics of the other items in the test. Finally, in order to obtain the number-correct score needed to predict item performance, the student has to take that item; the prediction would provide no increase in information or efficiency.

Avoiding these disadvantages, the IRT mathematical function, or *model*, defines a smooth S-shaped line, as in Figure 1(b). In place of the number-correct score, IRT models use a *trait* scale, also called the Θ (theta) scale, as the ability measure. This trait scale is designed to be stable when different items are in the test. The IRT model also has a small number of *item parameters* that take on different values for different items. The result is that the entire ICC of each item can be described with just a few numbers.

Having ICCs that are stable over groups of students is sometimes described as *person-free item measurement*. Being able to produce equated test scores from a variety of sets of items is sometimes called *item-free person measurement*. These catchy terms are exaggerations relative to what is found with real-life tests, but they are useful for conveying the ideals that motivate IRT. As described more fully in the Model Fit section, to the extent that IRT's simplifying assumptions are true, a model will be useful and make accurate predictions; to the extent that they are not true, use of a model can lead to inaccuracies.

FIGURE 1. *Item characteristic curves conditional on number-correct scores and IRT ability scores*



Survey of Educational Applications

When IRT is used appropriately, it can increase the efficiency, accuracy, or usefulness of a wide variety of measurement processes. Many of the following advantages can be obtained by one or another of the classical procedures, but the IRT models provide a unified framework and system that facilitates their accomplishments. More detailed information about some of these applications is presented in later sections.

Test Construction. An IRT model can be used to create a pool of items that have known statistical characteristics, including descriptions of how well each item is measuring students at each ability level. Measures of *differential item functioning* can be provided for use in evaluations of item bias. The psychometric properties of any test created from the pool can be readily predicted for different groups of students, even when those students have not taken that test. These properties include number-correct score means, standard deviations, and distributions, as well as reliabilities, standard errors of measurement, and item p -values. Thus the IRT models are ideal for computer-assisted item selection systems, which give the test constructor suggestions for items that meet various needs and instant feedback on the effects of alternative item selections.

Test Equating and Administration. Traditional equating procedures produce conversion tables that show how to translate a score on one test to a comparable score on another test. With traditional procedures, the intact tests are administered to students to collect the equating information. However, IRT procedures are more flexible, because the item, rather than the intact test, can be the unit of scaling or equating. Once item scaling has been successfully accomplished, the items can be selected for a variety of test configurations. The IRT model states how to aggregate the item information to get test information and how to produce equated ability scores for different tests.

This flexibility can be used in computerized *adaptive* or *tailored* testing, in which a student's response to each item is considered in choosing the next item to be administered. Adaptive testing chooses items to match the student's performance level and requires fewer items than an intact test to reach a desired level of score accuracy.

The use of IRT is well suited to matrix sampling, in which multiple test forms are created and administered to different students using a random sampling procedure. Matrix sampling is useful for obtaining group-level data on a broad sample of items in a specified content domain, while limiting the testing time required of each student.

Test Scoring and Interpretation. The user of any test score should know the amount of measurement error it is likely to contain. Classical test theory produces a single standard error of measurement (SEM) that ap-

plies to all scores obtained with a test. However, IRT goes beyond the classical approach to provide a different SEM for each score. For example, if a test emphasizes easy items, scores for low-ability students will be more accurate than those for high-ability students.

Indices are available that reflect the appropriateness of a test in measuring a given student. For example, if a student's score appears to have been influenced by substantial guessing or noncompletion of the test, the score can be flagged. In some cases, adjusted scores can be provided.

Testers can use IRT in test scoring to increase accuracy by taking into account the statistical characteristics of the particular items that the student answered correctly. Such scoring methods can be particularly helpful in increasing score accuracy for low-scoring students who have taken multiple-choice tests.

From a student's score on a subset of the items in an item pool, IRT can yield that student's probability of passing any of the other items in the pool. Thus scores can be referenced extensively to content, enhancing interpretation and instructional decisions.

Models and Their Parameters

Most IRT models used in practice utilize one characteristic of the student, usually called the *trait*, and are called *unidimensional* models. The numerical value of the trait reflects the level of the student's ability, achievement, skill, or other primary characteristic being measured by the test. The trait is alternatively called the latent trait, ability, theta (Θ), or scale score. In this article, the term *ability* is used.

These models also typically employ one, two, or three item characteristics, or *parameters*, that reflect differences among the items in their statistical attributes. Using ability, item parameter(s), and a mathematical function, the model describes $P_i(\Theta)$, the probability that a student with ability Θ will pass item i . The value of i ranges from 1 to n , the number of items in the test. By definition, the probability that the student will fail item i is $1 - P_i(\Theta)$. The three most commonly used models contain the same mathematical function, the *cumulative logistic*.

Three-Parameter Model. The three-parameter model is the most general of those in common use. It states that

$$P_i(\Theta) = c_i + \frac{1 - c_i}{1 + \exp[-Da_i(\Theta - b_i)]} \quad (1)$$

The term $\exp[y]$ means "take the constant e (2.718...) to the power y ." The term D is a constant that is usually taken as 1.7. Students differ in terms of their Θ values, with higher values indicating a higher chance of cor-

rectly answering a specific item and all the other items in the test. As discussed in the Scale Units section, different Θ scales can be used interchangeably to a certain extent; the examples described here have Θ values from about 100 to 400. Items differ in terms of the numerical values of their a_i , b_i , and c_i parameters.

The c_i parameter is the probability that a student with very low ability will answer the item correctly. For multiple-choice items, where students can choose the correct answer by guessing, c_i values are usually close to the reciprocal of the number of answer choices; for example, the c_i for a four-choice item will usually be in the neighborhood of .25. The actual value of c_i will be influenced by how attractive the wrong answer choices are to very low ability students. Thus, although the c_i value can be influenced by nonrandom as well as random factors, it typically is called the *guessing parameter*. The c_i value is also called the *item lower asymptote* because the ICC does not reach or get lower than c_i no matter how low the student's ability. Figure 2(a) displays the ICCs of two items that differ only in terms of their c_i values, with $c_1 = .25$ and $c_2 = .15$.

The b_i parameter is called the *item difficulty*. The higher the b_i value, the more difficult the item is and the less likely that a student will answer it correctly. Figure 2(b) displays the ICCs for two items that differ only in terms of their b_i values, with $b_1 = 250$ and $b_2 = 300$. Item 1 is the easier item, and a student at any ability level has a higher probability of getting Item 1 correct than Item 2. Note that b_i and Θ are in the same units and that when a student's ability equals an item's b_i value, the student's probability of getting the item correct is halfway between c_i and 1.0. The b_i is sometimes called the *item location*, because it describes the location of the ICC on the Θ scale. The term item location also is used so that the IRT parameter will not be confused with the traditional item difficulty, the p -value. As items become more difficult, b_i values get higher while p -values get lower.

Figure 2(b) shows that the probability of correctly answering each item is changing most rapidly when $\Theta = b_i$. The third item parameter, a_i , is a function of how steep the ICC is at $\Theta = b_i$ and is called the *item discrimination*. The higher the a_i value, the more the item distinguishes or discriminates among students whose abilities are about the same as the item's difficulty. The items in Figure 2(c) differ only in terms of their item discriminations, with a_2 being twice as large as a_1 .

Figure 2(d) contains ICCs for items with several different combinations of item parameters. A wide variety of ICC shapes can be described with the three-parameter model.

Two-Parameter Model. A special case of the three-parameter model, the two-parameter model, states that $c_i = 0$ for all items. That is, students with very low ability are not expected to answer any items correctly. This assumption would be appropriate for completion items

or in cases where students cannot or do not guess the correct answer. The two parameters of this model are item discrimination (a_i) and difficulty (b_i), each of which can vary over items. Figure 2(e) shows ICCs for three items that are consistent with the two-parameter model.

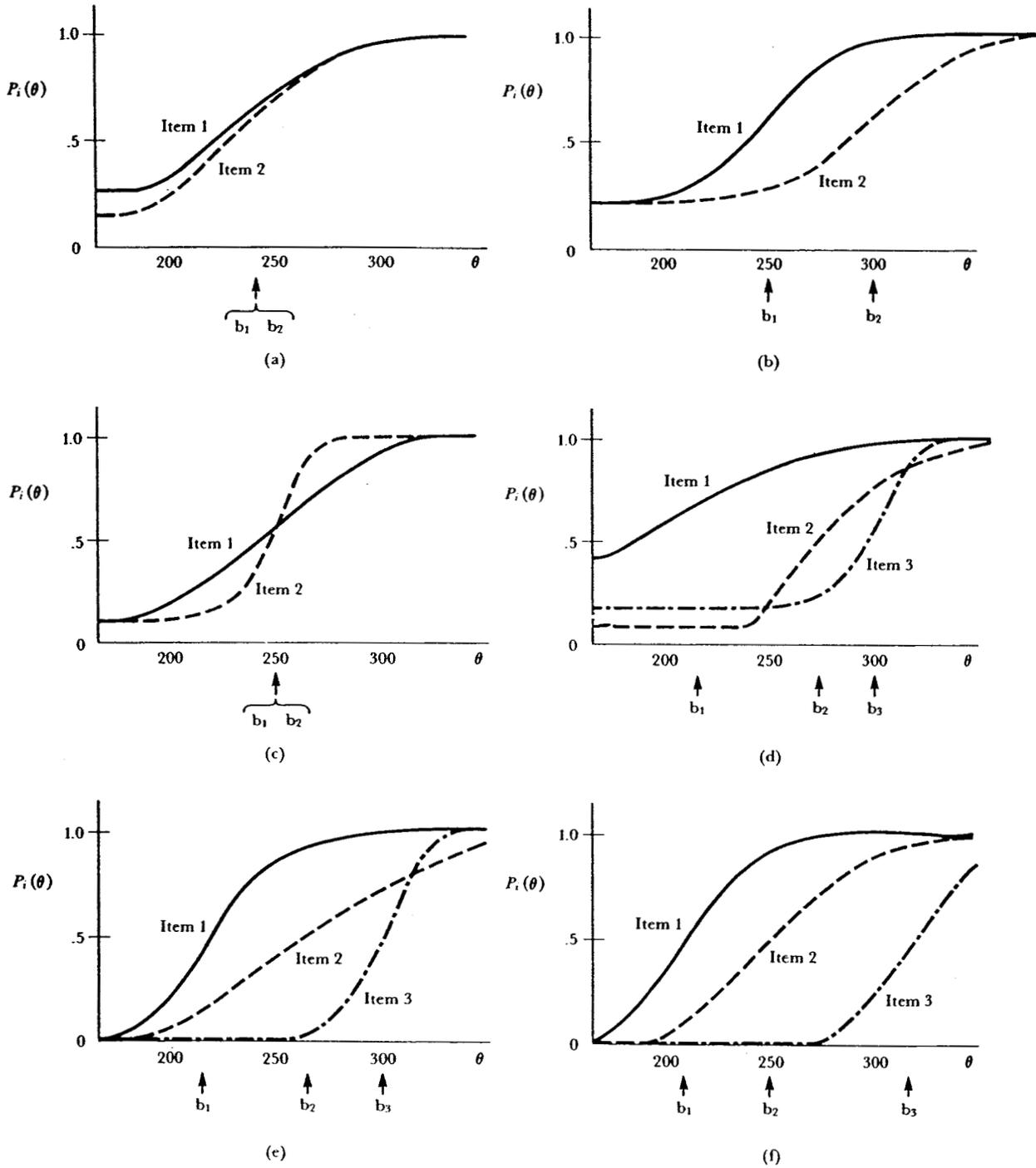
One-Parameter Model. The most restrictive of the models in common use, the one-parameter model, is a special case of the three- or two-parameter model in which $c_i = 0$ for all items, and the item discriminations are all equal. The constant value used for the item discriminations is arbitrary, but typically a_i is taken as $1/D$, with the result that the Da_i term no longer appears in the item characteristic function. In the one-parameter model, items can vary with respect to one characteristic only: their difficulty (b_i). Figure 2(f) shows ICCs for three items that are consistent with the one-parameter model.

The one-parameter model is commonly called the *Rasch model*. Georg Rasch (1960) conducted extensive research on this model, examining it not as a special case of the three- or two-parameter model but as the only model that produced certain desired measurement characteristics.

Local Independence. If an IRT model is true, the student's ability and the items' parameters incorporate all the important information about the student's performance on the items. This model property implies that performance on different items is independent, that is, *conditional on the student's ability*. This *local independence* means that the probability that a student will answer correctly any two items is the product of the probabilities that the student will answer correctly each separate item, and the psychometric contribution of an item to a test can be evaluated without knowledge of the other items in the test. Local independence does not occur when a test is speeded or when the stem or answer to one item gives information that can be used in answering another.

Scale Units. One property that all the IRT models share is that the scale for the ability and item parameters can be linearly transformed without changing the models' predictions. A linear transformation involves multiplying by one constant and/or adding another constant. Most (but not all) of the computer programs that estimate item parameters and abilities define the scale so that the abilities in that analysis have (approximately) a mean of 0 and standard deviation of 1. For educational applications this scale commonly is transformed to avoid negative numbers—for example, to have a mean of 300 and a standard deviation of 50. Such a linear transformation does not affect the models' predictions, so long as an appropriate change is made to the item parameters. The $Da_i(\Theta - b_i)$ term is the only part of the model influenced by the choice of scale. Imagine that a new scale is defined using $\Theta^* = 50 \cdot \Theta + 300$, $b_i^* = 50 \cdot b_i + 300$, and $a_i^* = a_i/50$. Then, $Da_i^*(\Theta^* - b_i^*) = Da_i(\Theta - b_i)$, and for every student and item the new scale makes the same predictions as the old scale.

FIGURE 2. Examples of item characteristic curves



The fact that the predictions based on these models are not affected by linear transformations makes them consistent with traditional criteria for producing *equal-interval* scales, in which an increase of one score unit has the same meaning anywhere on the ability scale. Numerous non-IRT procedures also exist for producing equal-interval scales. Although these procedures all produce putative equal units, their results can differ systematically, and no objective criterion exists for deciding which scale is the "right" one. Scaling units can be important in education when they are used to draw conclusions about academic growth. In examining growth, regardless of whether an IRT or non-IRT scale is involved, the prudent tester uses analysis techniques whose conclusions are not affected by the choice of scale. This caution is particularly important when the amounts of growth for students at very different achievement or ability levels are to be compared (Braun, 1988; Yen, 1986).

Choosing a Model

Opinions vary about how to choose which, if any, of the IRT models to use, and these opinions can be described as falling along a continuum. At one end are those who "choose the model to fit the data," and at the other end are those who "choose the data to fit the model."

The *choose the model* approach finds a model that is sufficiently general to fit the item data with desired accuracy. The less restrictive a model is, the more likely it will be found to match real-life student performance. For example, because the two- and three-parameter models allow items to vary in terms of item discrimination and the one-parameter model does not, the two- and three-parameter models are more accurate when items in fact vary in terms of their discriminations. In choosing a model to optimize fit, the tester would be led to the more elaborate model. In the extreme, one would theoretically be led to use no model at all but to return to a complete description of the data at hand, as in Figure 1(a). Although such an extreme approach would be optimally accurate, it would also be inefficient, because it would not take advantage of the consistencies and generalities that do occur in item data.

From the *choose the data* perspective, the measurement properties that are desired are identified and the model(s) that produce these properties determined. Then, only those items that are consistent with the ideal model are used. To the extent that a model's requirements are unnecessarily restrictive, selection of fitting items can have an unnecessary negative impact on the content or construct validity of the test. In the extreme, this approach leads to abandoning measurement, because in real life items conforming to the ideal cannot be produced.

Proponents and detractors of various IRT models can be described as falling somewhere on this continuum. Heated disagreements have arisen about how model fit should be evaluated and whether the fit statistics used have sufficient statistical power to detect important instances of misfit. In general, model fit is not an all-or-none condition, and a model's predictions can be judged to be sufficiently accurate for one application but not another. These issues are discussed further in the Model Fit section.

Selected Educational Applications

Predicting Group-Dependent Test Performance. Determining how an item or a test works for any group of students begins with consideration of the abilities of the students in that group. Let Θ_k be the ability of student k , and k ranges from 1 to N , the number of students. For that group, the p -value, or expected proportion passing item i , is

$$p_i = \frac{1}{N} \sum_{k=1}^N P_i(\theta_k). \quad (2)$$

An example of the use of this procedure is with a pool of calibrated items; Θ_k is estimated by having students take a subset of the items in the pool. Equation 2 is used to estimate how these students would have done on any of the items in the pool they did not take. These predictions are combined for sets of items to obtain predictions of a wide variety of test statistics, such as number-correct score means, standard deviations, and reliabilities.

Standard Error of Measurement. In classical test theory, the accuracy of scores is expressed in terms of test reliability, $r_{xx'}$, and inaccuracy in terms of the standard error of measurement. In classical theory the SEM equals $S_x \sqrt{1 - r_{xx'}}$ where S_x is the standard deviation of observed test scores. The classical SEM is a single, group-dependent statistic that communicates an average amount of score inaccuracy.

Going beyond the single SEM for a test, IRT describes the SEMs that a test has at each ability level and the independent contribution that each item makes to those SEMs. At a given ability estimate, $SEM(\hat{\Theta}) = 1/\sqrt{I(\hat{\Theta})}$ where $I(\hat{\Theta})$ is called the *test information*. As test information increases, the SEM decreases. When optimal procedures are used in scoring a test (as described in the Test Scoring section), the test information is maximized and equals the sum of the *item information functions*, $I_i(\hat{\Theta})$. All else being equal, the more items there are in a test, the greater is the test information obtained and the lower is the SEM function.

Figure 3(a) shows the information functions of the items in Figure 2(d). (The formula for $I_i(\hat{\Theta})$ can be found in most IRT texts—for example, Lord, 1980.) Items provide the most information near their locations, and items

with higher discriminations and lower guessing parameters provide more information. The information function for a test composed of those three items is also shown in Figure 3(a) and its SEM function in Figure 3(b). The SEM is lowest near scores of 300, because the items that measure best are located near that score.

Test users sometimes need to choose among several available test forms or levels. Comparison of SEM functions or information functions, called analysis of *relative efficiency*, can give precise information about which test would be more accurate for a particular group of students.

Item Pools and Test Construction. As described more fully in the Item Parameter Estimation section, students' responses to a set of items are used to estimate item parameters, or *calibrate* the items. Frequently, the test developer wants to construct a pool of items, all of which are calibrated to the same scale. This pool would contain more items than would be prudent or feasible to administer to one student. Multiple test forms are constructed and administered to different groups of students, and then one of several possible equating procedures is used to align or equate the results from the different forms and place them on the same scale.

In some cases the groups of students that are given different forms are carefully selected to be equivalent, which becomes the basis of equating the results in the different forms. In another equating technique, a subset of the items are the same across forms; these common items are called *anchor* items. By aligning the parameters of the anchor items, all the items in the different forms are calibrated to the same scale (Stocking & Lord, 1983).

After the pool has been constructed, items can be readily selected to create tests for specific applications. For example, different tests can be created to match the performance levels of students at different grades. Identifying the best items for determining minimum competency or for selecting students for advanced instruction is a straightforward procedure. Extensive psychometric descriptions can be produced to describe the characteristics of any test selected (Green, Yen, & Burket, 1989).

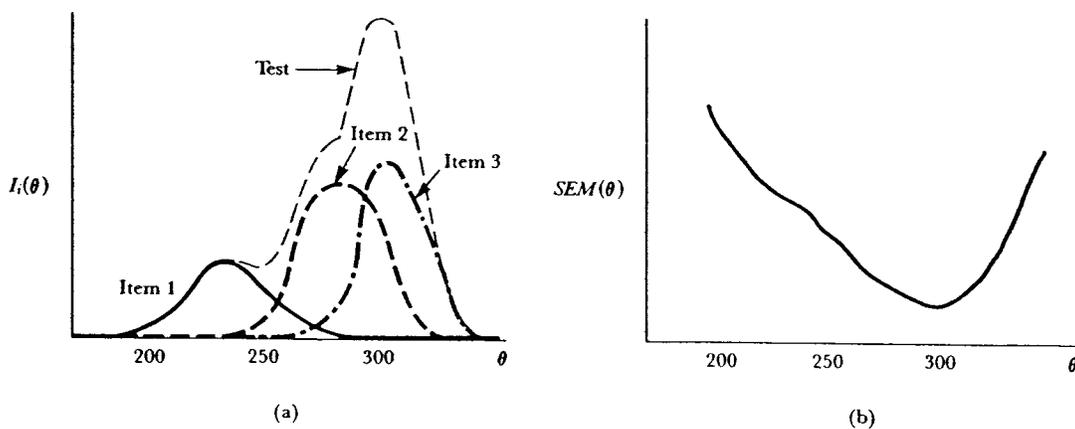
Test Scoring. When item parameters have been obtained, they can be used to produce equated ability estimates for different tests. There are (at least) two types of performance data that can be used to score a test to obtain an ability estimate: number-correct scores and patterns of item responses. Given the performance data, a well-established statistical procedure called *maximum likelihood* is commonly used to estimate the ability for each student.

For the one-parameter model, the maximum likelihood ability estimate is that which has an expected number-correct score equal to the student's observed number-correct score. That is, if X_k is the number-correct score for student k , that student's ability estimate is the $\hat{\theta}_k$ that makes

$$X_k = \sum_{i=1}^n P_i(\hat{\theta}_k). \quad (3)$$

Equation 3 transforms a number-correct score on a particular test to an estimated ability. It can be applied to any test composed of items that have been successfully calibrated to the same ability scale. For every test, a student's estimated ability, $\hat{\theta}_k$, is expected to be the same, or *equated*, except for random variation due to measurement error; the amount of measurement error expected is described by the SEM function. By transforming number-correct scores on different tests to equated ability estimates, Equation 3 also determines the number-correct score on one test that is equivalent to any given number-correct score on another.

FIGURE 3. Item information, test information, and test SEM functions based on the items in Figure 2(d)



For every IRT model, Equation 3 can be used to get a maximum likelihood ability estimate. If the two- or three-parameter model is used, it is possible to go further to increase the accuracy of the ability estimate by considering which particular items the student answered correctly (that is, considering the *item response pattern*). Compared to number-correct scoring, pattern scoring raises some students' ability estimates and lowers others', producing no change in the average ability estimate for groups of students. Pattern scoring is optimal, because it takes into account all the information available in the item responses and has lower SEM values than number-correct scoring.

Score Interpretation. By placing item difficulties on the same scale as a student's ability, IRT can be used to create informative score reports. The student's ability can be estimated with a subset of the items in a pool. This ability estimate can then be used to predict the probability that the student would have passed items in the pool that were not taken. Such information is useful in relating a student's score to concrete performance descriptions. Examples of such score interpretations are provided by Connolly, Nachtman, and Pritchett (1976) and Mullis and Jenkins (1990).

Item Bias Information. A major concern in the evaluation and construction of educational tests is that they have as little ethnic and gender bias as possible. There are many definitions of *bias*, and IRT is particularly well suited to providing information relevant to one of them.

An item shows *differential item functioning* (DIF) if the conditional probability of getting the item correct—based on the student's ability, $P_i(\theta)$ —differs systematically for students who are members of different groups of interest. In other words, ICCs that differ over groups are evidence of DIF.

Note that DIF is *conditional* on ability. The fact that two groups differ in their p -values is not an indication of DIF, because p -values are influenced by differences in the distributions of ability as well as by differences in the conditional relationship between ability and item performance. Moreover, DIF procedures compare items relative to each other; constant effects that influence all items cannot be detected with DIF analyses.

Various procedures are available for evaluating DIF with IRT. Discussions of these procedures are presented in Hambleton (1989), Cole and Moss (1989), and Hulin, Drasgow, and Parsons (1983).

Item Parameter Estimation

To obtain item parameter estimates, or to *calibrate* a set of items, it is necessary to have the responses of a group of students to these items. The procedure used to estimate the parameters is iterative. They typically begin with some starting estimates for abilities and then—holding these values as fixed or known—estimate item

parameters. Then, treating the item parameters as known, the abilities are reestimated. This process continues until satisfactory convergence in the estimates is obtained. Because of the complexity of the estimation process, it is almost always used with a computer program run on a mainframe or microcomputer.

Simulation studies have shown that programs are available that can accurately estimate the parameters and abilities for the one-, two-, and three-parameter models. The more item parameters there are in the model, the greater is the number of students needed to obtain accurate parameter estimates. There are no hard and fast rules about the minimum number of students or items needed for useful estimates to be produced, but only rarely would a test and group of students produce useful estimates with fewer than, say, 15 to 20 items and fewer than 200, 400, or 600 students for the one-, two-, or three-parameter models, respectively. The level of accuracy needed for the particular application needs to be considered. For example, much more accuracy is needed for equating tests used to award a medical license than is needed for an item tryout whose purpose is to get a rough idea of the grade level at which items measure best. Formulas for the standard errors of maximum likelihood item parameter estimates are available (Lord, 1980). Detailed comparisons of the characteristics and performance of the estimation programs are available in Hambleton (1989), Hsu and Yu (1989), Lord (1986), and Mislevy and Stocking (1989).

Model Fit

The usefulness of an IRT model is its ability to describe and predict students' performance on items. If an IRT model could be shown to be completely accurate, all of its predictions would be true. However, verifying every prediction is not possible without collecting the empirical data that were the object of the prediction, thereby eliminating the efficiency that was the motivation for using the model. Thus those using IRT models typically describe the level of accuracy of the predictions in which they have special interest and generalize these findings to similar situations. Many procedures are available to examine model fit, and three basic types are described here: (a) indirect analyses; (b) residuals; and (c) measurement characteristics.

Indirect Analyses. Indirect studies typically are done before users make the intellectual and financial investment in the knowledge and computer programs required to actually analyze their data with one or more of the models. These analyses investigate whether existing tests or data might be clearly inconsistent with the models. An analysis of test speededness may be conducted; if substantial numbers of students get items wrong because they do not finish the test, it is unlikely that an IRT model will be suitable. A rational or empir-

ical analysis may be made of the chance that students will try or succeed in guessing the correct answer: the greater this chance, the less likely that the two-parameter or Rasch models will be suitable.

Residuals. After parameters and abilities have been estimated, fit analyses can be conducted that compare observed and predicted ICCs. The observed performance for student k on item i takes on the value 0 for an incorrect answer and 1 for a correct answer. The predicted performance for that student on that item is $\hat{P}_k(\hat{\theta}_k)$. The difference between the observed and predicted values is the *residual*, or error of prediction.

A variety of procedures are available for examining item residuals (Hambleton, 1989; Ludlow, 1986; Wright & Stone, 1979). Residuals can also be examined for a single student to determine whether the student's performance indicates guessing, inconsistency, or some other characteristic of interest (Hulin, Drasgow, & Parsons, 1983); this examination is called *appropriateness measurement* or *person fit*.

Measurement Characteristics. Analyses of residuals often do not reveal the practical consequences of misfit, and further studies are made with that focus. For example, predictions of number-correct score means, standard deviations, distributions, and SEM functions can be made and checked.

Of particular interest have been examinations of the degree to which item parameter estimates are *person-free* and ability estimates are *item free*. The latter is an examination of test equating and is the IRT measurement characteristic that has been most thoroughly researched. In fact, research into IRT equating has revived interest in examining the adequacy of other equating methods, as well. Results of such studies have varied greatly, though some generalizable conclusions can be drawn.

1. When an item appears in different test forms, the more similar its surrounding context, the more stable the item's statistical qualities will be. If fatigue is a factor, items can be more difficult if they appear at the end of a test.

2. The more similar students' test-relevant experiences are, the more likely that item parameters will be stable for them. However, DIF analyses indicate that experts cannot straightforwardly judge from item content which items will be most stable over groups.

3. The more equal the difficulty of test forms being equated, the better the IRT (as well as the other equating methods) works. Because the one-parameter model does not adjust for guessing, its use may present particular problems in equating multiple-choice tests that vary substantially in difficulty.

4. The more homogeneous the content of a test, the more likely an equating based on IRT (or any other single-score method) will hold up over different populations of students. Test forms, including customized tests, must

be matched in terms of content coverage if they are to produce comparable scores.

5. The numbers of items and students must be large enough to be suitable for the equating methods being used, with more items and students being required for procedures that describe more details of the test performance.

One educational equating issue deserves special attention: IRT has made customizing tests easier, that is, systematically selecting items for particular applications. In some cases local educational agencies have developed their own tests and then used IRT to equate their tests to nationally normed tests. In other cases, test publishers have provided customized versions of their normed tests to such agencies, linking the customized test to national norms. In evaluating the appropriateness of these procedures, the principles just described apply. In particular, the more the customized test narrows the content or changes its emphasis from that in the normed test, the less accurate its normative scores will be. There are legitimate reasons for customizing tests, such as to integrate norm-referenced measurement with other parts of a test program in order to improve efficiency, and valid normative information can be obtained from customized tests. However, customization needs to be done with care (Yen, Green, & Burket, 1987).

Future Developments

Interest in IRT models likely will continue and expand. For example, increasing attention is being focused on the fact that test information can be invalidated if test security is compromised or if students focus on learning specific test items rather than the skills the test is attempting to measure. These concerns can be addressed efficiently through the use of IRT to create item pools from which multiple test forms are created. Matrix sampling using IRT also addresses those issues, as well as the issue of decreasing the amount of testing required of each student while maintaining broad content coverage (Bock, Mislevy, & Woodson, 1982). Content referencing of scores is becoming more widely used (Mullis & Jenkins, 1990). Increased use of microcomputers in education will increase the availability of IRT technology, including computerized adaptive testing (Bunderson, Inouye, & Olsen, 1989; Wainer, 1990; Weiss, 1983).

Item response theory models that are more elaborate than those presented here are likely to receive increasing attention. Three examples are multidimensional models involving more than one ability (Reckase, Ackerman, & Carlson, 1988); models that can be applied to ratings, such as those that result from the scoring of writing samples and other performance assessments (Thissen & Steinberg, 1986; Wright & Masters, 1982); and models especially designed to relate item performance to the cognitive demands of the items (Embretson, 1985; Sheehan & Mislevy, 1990).

Sources of Further Information

The *Journal of Educational Measurement* (Summer 1977) contains six articles introducing the basic theory and potential application of IRT (under the then-current name *latent trait theory*). McKinley (1989) provides an overview of IRT with a minimum of equations. The measurement textbooks by Allen and Yen (1979) and Crocker and Algina (1986) have some introductory IRT information. Baker's (1985) IRT introduction includes a micro-computer program that permits interactions with the models. Wright and Stone (1979) describe the philosophy and application of the Rasch model.

Familiarity with measurement and statistical concepts is assumed in the thorough IRT survey in Hambleton (1989) and in the collection of 13 applications in Hambleton (1983). *Applied Psychological Measurement* (December 1986) presents nine articles on item banking. Hambleton and Swaminathan (1985) and Hulin, Drasgow, and Parsons (1983) provide IRT textbooks, leading from an introductory to a more advanced level. At the advanced level Birnbaum (1968) describes fundamental concepts, and Lord (1980) presents an elegant and authoritative treatment of IRT.

Journals that publish most of the original articles dealing with IRT are *Applied Psychological Measurement*, *Journal of Educational Measurement*, *Journal of Educational Statistics*, and *Psychometrika*.

Wendy M. Yen

See also Achievement Testing; Item Analysis; Norms and Scales; Reliability of Measurement; Test Construction; Testing Technology.

REFERENCES

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Baker, F. B. (1985). *The basics of item response theory*. Portsmouth, NH: Heinemann.
- 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. 397-472). Reading, MA: Addison Wesley.
- Bock, R. D., Mislevy, R., & Woodson, C. (1982). The next stage in educational measurement. *Educational Researcher*, 11(3), 4-11.
- Braun, H. I. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement*, 25, 171-191.
- Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 367-407). New York: American Council on Education/Macmillan.
- Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: American Council on Education/Macmillan.
- Connolly, A. J., Nachtman, W., & Pritchett, E. M. (1976). *Math diagnostic arithmetic test*. Circle Pines, MN: American Guidance Service.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
- Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.
- Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. *Applied Measurement in Education*, 2, 297-312.
- Hambleton, R. K. (Ed.). (1983). *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.
- 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: American Council on Education/Macmillan.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.
- Hsu, T., & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice*, 8(3), 21-28.
- Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, 23, 157-162.
- Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement*, 10, 217-229.
- McKinley, R. L. (1989). An introduction to item response theory. *Measurement and Evaluation in Counseling and Development*, 22, 37-57.
- Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, 13, 57-75.
- Mullis, I. V. S., & Jenkins, L. B. (1990). *The reading report card, 1971-1988*. Princeton, NJ: National Assessment of Educational Progress.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut (Chicago: University of Chicago Press, 1980).
- Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193-203.
- Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, 27, 255-272.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567-577.
- Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA.

Yen, W. M. (1986). The choice of scale for educational measure-

ment: An IRT perspective. *Journal of Educational Measurement*, 23, 299-325.

Yen, W. M., Green, D. R., & Burket, C. R. (1987). Valid normative information from customized achievement tests. *Educational Measurement: Issues and Practice*, 6(1), 7-13.