Item Response Theory

Robert J. Harvey
Virginia Polytechnic Institute and State University

Allen L. Hammer
Consulting Psychologists Press

Item response theory (IRT) seeks to model the way in which latent psychological constructs manifest themselves in terms of observable item responses; this information is useful when developing, evaluating, and scoring tests. After providing an overview of the most popular IRT models (i.e., those applicable to dichotomously keyed items) and contrasting them with the techniques used in classical test theory (CTT), the authors illustrate the application of IRT using data from the recently revised Myers-Briggs Type Indicator. These results highlight a number of IRT's advantages, including (a) detailed descriptions of the performance of individual items, (b) indices of item- and scale-level precision that are free to vary across the full range of possible scores, (c) assessments of item- and test-level bias with respect to demographic subgroups, (d) measures of response-profile quality, and (e) computer-adaptive testing, which can dramatically reduce testing time.

Although item response theory (IRT) methods have been in existence for more than half a century (e.g., Lord, 1952; Tucker, 1946), only recently have they begun to achieve widespread popularity in psychological assessment, especially outside the realm of large-scale, standardized aptitude and achievement testing. One very practical reason for this belated popularity is the fact that IRT techniques tend to be far more computationally demanding than methods of test construction and scoring that are based on classical test theory (CTT); prior to the widespread availability of efficient computer software (e.g., Mislevy & Bock, 1983) and affordable computer hardware, IRT methods were simply too difficult and expensive for most test users and developers to implement. Indeed, some of the most useful applications of IRT technology—such as computer-adaptive testing (CAT) (e.g., Sands, Waters, & McBride, 1997)—became practical only as a result of the dramatic improvements in computer price and performance that have occurred over the past 10 to 15 years.

Initially, IRT methods were developed primarily for use with standardized achievement and aptitude tests composed of multiple-choice items scored in a "right-wrong" format, such as the Scholastic Aptitude Test (e.g., Lord, 1968). Although instruments of this type are indeed used by counseling

Corespondence should be addressed to Robert J. Harvey, Department of Psychology, Virginia Polytechnic Institute & State University, Blacksburg, VA 24061-0436; e-mail rj@pstc.com.

THE COUNSELING PSYCHOLOGIST, Vol. 27 No. 3, May 1999 353-383

psychologists, measures of personality, attitudes, and interests are also critically important. Fortunately, IRT methods are not confined to traditional ability and achievement tests, and they are increasingly being applied to personality, attitude, and similar inventories containing items that are scored in a dichotomous fashion, such as checklists and inventory-type items that can be "keyed" in a given direction (e.g., Brown & Harvey, 1998; Drasgow & Hulin, 1990; Harvey & Thomas, 1996). Recently, increased attention has also been devoted to IRT models that are capable of analyzing items that are rated using either ordered-category scales (e.g., Likert-type scales) or unordered, nominal scales (e.g., Bock, 1972; Samejima, 1979; Thissen & Steinberg, 1985); the addition of these polytomous models renders the IRT approach applicable to virtually any type of standardized psychological assessment instrument.

As will be described below, IRT offers many important advantages over CTT-based methods of test development and scoring; indeed, now that the primary technological obstacle to its widespread use (i.e., the availability of inexpensive, powerful personal computers) has been removed, it is not unreasonable to predict that IRT-based methods—also referred to as "modern test theory" (e.g., Embretson, 1996)—will largely replace CTT-based methods over the coming years. Consequently, it is important that counseling psychologists become familiar with both the theoretical and practical aspects of IRT so that they will be prepared to interpret inventory results for their clients, as well as to select, evaluate, and develop new tests that use IRT technology. Consistent with its origins in tests of educational achievement and aptitude, IRT methods are already well known among educational researchers—especially the 1-parameter, or Rasch, model (e.g., Wright, 1977). IRT has also achieved wide use among industrial and organizational psychologists (e.g., Drasgow & Hulin, 1990), in part due to its ability to quantify the degree to which tests exhibit consistent bias with respect to race, sex, age, or other demographic factors.

Our goal in this article is to provide an overview of the most popular IRT models and then illustrate the practical application of IRT methods using the recently revised Myers-Briggs Type Indicator (MBTI) (Myers, McCaulley, Quenk, & Hammer, 1998). The MBTI results highlight a number of the ways in which IRT provides a richer view of item- and test-level performance than is possible using CTT methods; in addition, they underscore the important fact that IRT methods are not limited to traditional ability and aptitude tests, providing the same benefits for personality, interest, and other inventories that do not employ right-wrong items and response formats.

OVERVIEW: IRT MODELS AND METHODS

IRT models have been developed to deal with responses to items that are scored in either a dichotomous (i.e., only two possible scored responses exist, such as true-false, correct-incorrect, endorsed–not endorsed, etc.) or a polychotomous or polytomous (the former term being used in earlier IRT literature, with the latter supplanting it in more recent years) fashion (i.e., more than two scored values are possible, such as Likert-type attitude or opinion-survey items). With regard to the former category, it is important to emphasize that the IRT models for dichotomous items are not restricted to two-alternative multiple-choice formats; that is, they can be applied to multiple-choice items that possess any desired number of response alternatives and even to non-multiple-choice, free-response items. In short, the primary requirement is that each person's item response has the ability to be scored to produce a dichotomy, not that the item response itself was dichotomous, or that the item was phrased in a right-wrong fashion.

A consideration of all of the IRT models that have been advanced to date is well beyond the scope of this article; instead, we will limit our coverage to IRT models for dichotomously scored items, given that (a) these IRT models have received considerably more research attention, and practical use, than models for polytomous data (although this situation may change in the future, as polytomous models become more fully developed, and they see wider application); (b) an understanding of the dichotomous IRT models is effectively a prerequisite for dealing with the more complex polytomous models; and (c) dichotomous IRT models are applicable to a broad range of assessment instruments. For readers interested in more detailed treatments of IRT in general, or polytomous IRT models in particular, a number of sources are available for further reading (e.g., Drasgow & Hulin, 1990; Hambleton, Swaminathan, & Rogers, 1991; Hulin, Drasgow, & Parsons, 1983; Lord, 1980; Lord & Novick, 1968; Samejima, 1979; Sands et al., 1997); in particular, the van der Linden and Hambleton (1997) text provides a comprehensive overview of a number of more advanced IRT models and topics (e.g., partial-credit models, models allowing multiple response attempts, as well as nonparametric, nonmonotone, and multidimensional models).

Assumptions and Terminology

Unidimensionality. Traditionally, IRT models have been based on the assumption that the item pool being analyzed is effectively unidimensional,

although some attempts to develop multidimensional IRT models have been made (e.g., Muthen, 1984). Given that IRT techniques are typically applied to instruments whose dimensional structures have already received significant empirical study (e.g., using exploratory or confirmatory factor analytic methods), the assumption of unidimensionality does not typically represent an undue practical restriction. That is, for instruments composed of multiple subtests or scales, each subtest can simply be analyzed separately using a unidimensional IRT model (as was done when IRT was used to develop the most recent revision of the MBTI, discussed below).

Of course, in practice, no scale composed of a reasonable number of items will ever be perfectly unidimensional. Fortunately, research designed to assess the impact of violations of the unidimensionality assumption (e.g., Drasgow & Parsons, 1983; Hulin et al., 1983) has suggested that the unidimensional IRT models are relatively robust with respect to moderate violations of strict unidimensionality and that the most important issue concerns the relative degree to which the item pool is dominated by a single latent trait.

The latent trait. The unobserved characteristic that is presumed to be responsible for the observed responses that are made to the test's items is denoted *theta* (θ); it is analogous to the "true score" in CTT. For convenience, θ is assumed to be scaled as a *z* score, although the θ metric can be transformed to any desired unit size and origin. In unidimensional IRT models, the observed responses to a test item are assumed to be determined by the joint action of θ and the characteristics of the item in question (e.g., difficulty, discrimination).

Homogeneous subpopulation (HSP). The concept of the HSP is important for understanding a number of issues in IRT; fortunately, it is a simple one. Rather than being an assumption of IRT models per se, an HSP is instead simply a collection of individuals who are homogeneous with respect to their scores on the underlying construct (θ) being assessed. For example, in a large administration of a test given to 10,000 individuals, we might find 100 who score 1.6 *z* units above the mean. In most cases, the HSP is defined using the "true" θ score; if HSPs are formed in practice (e.g., when computing empirical item characteristic curves; see below), the estimated θ score must be used instead.

Probability of item endorsement (PIE), or probability of a correct response (PCR). IRT is an item-focused approach, and consequently, the most basic data used in IRT are the responses to individual test items. In optimal-performance tests, these are the scored right-wrong responses; for inventory-format tests, some form of keying system is applied to convert the

item responses to dichotomous responses (e.g., for a scale on a personality test designed to measure introversion-extraversion, the "introvert" response or responses to each item could be selected as the keyed response). For any given sample of examinees, the PCR/PIE is defined as the proportion of respondents, in each HSP of interest, giving the correct (or keyed, in the case of an inventory-format test) response to the item.

Item characteristic curve (ICC), or item response function (IRF). One of the most important relations in IRT is the one that exists between the underlying construct of interest (θ) and the response to each test item; indeed, the primary differences between the various IRT models concern the form of the causal relationship that is presumed to exist between θ and the observed item response. The ICC or IRF (these terms are used interchangeably in the literature; in this article, we will use *ICC* to denote this functional relationship) is a two-dimensional scatterplot of θ (*x*-axis) by item-response probability (PCR or PIE), depicting the item response that would be expected from an HSP located at any given point on the underlying construct; Figure 1 presents a sample ICC.

The dark line in Figure 1 depicts the ICC for a moderately "difficult" item (if seen in a right-wrong test), the horizontal reference line corresponds to a 50% correct (endorsed) response level, and the vertical reference line locates the HSP that would be expected to get this item right 50% of the time (or, for a non-right-wrong item, to endorse the item in the keyed direction 50% of the time). In this example, the group of respondents located at θ = 1.0 (i.e., the 84th percentile in a normal distribution) would be expected to exhibit a .50 PCR level for this item; in contrast, in the HSP located at θ = 0 (i.e., the mean, in a *z*-score metric), we would expect only about 1 in 5 of the respondents to get this item correct (or endorse it in the keyed direction).

Of course, we do not mean to imply by the above example that IRT models necessarily assume that the θ distribution follows any particular form (e.g., a normal distribution); on the contrary, when using Bayes modal or similar methods to estimate θ scores (e.g., Bock & Mislevy, 1982), researchers are free to assume virtually any given shape for the population θ distribution. Likewise, when estimating the unknown item parameters of the IRT models, there is no need to make rigid assumptions regarding the form of the θ distribution (e.g., Bock & Aitkin, 1981).

IRT Models

All unidimensional IRT models share the assumption that a single underlying latent construct (θ) is the primary causal determinant of the observed responses to each of the test's items. They differ with respect to the way in

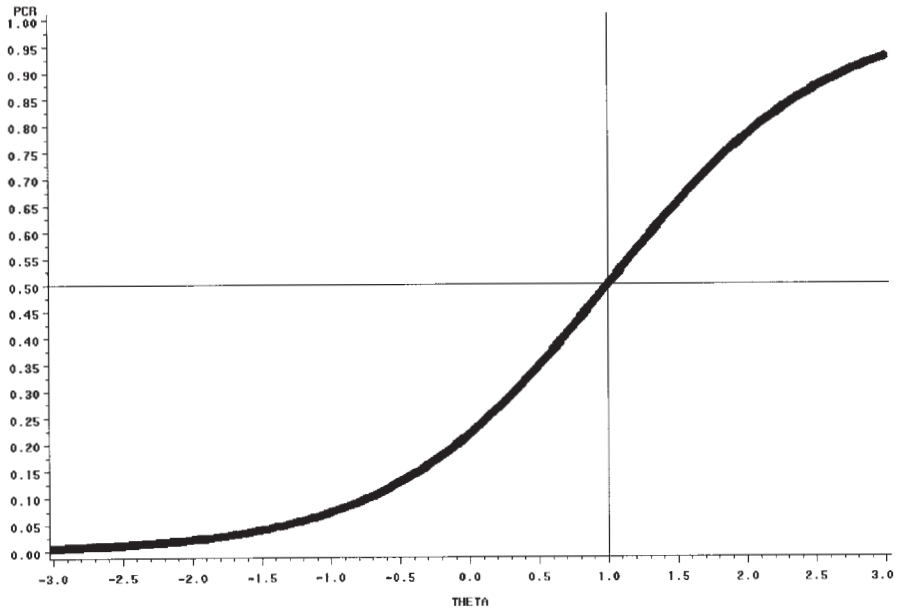

Figure 1. Item characteristic curve for a hypothetical test item.

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the observed probability of a correct response, in the case of a right-wrong item, or the probability of a response in the keyed direction. This item would be a moderately difficult one, with a θ score of 1.0 *z* units above the mean being required to achieve a 50% likelihood of producing a correct (or keyed) response. Item parameters are $a = 0.75$, $b = 1.0$, and $c = 0$.

which θ is presumed to cause the item response; the three most popular IRT models for item responses that can be expressed dichotomously are discussed below. In particular, these models differ with respect to the number of parameters they require in order to model the responses to each test item. In all cases, the item parameter or parameters effectively define the form of the causal relation that exists between θ and the observed item response; that is, the relation between θ and PCR/PIE is typically assumed to vary across the possible range of θ scores as a function of the item parameter or parameters.

1-parameter logistic model. The 1-parameter logistic, or Rasch (e.g., Wright, 1977), model is one of the simplest IRT models; as its name implies, it assumes that only a single item parameter is required to represent the item response process. In the 1-parameter logistic model, this item parameter is termed *difficulty* (abbreviated *b* in most of the IRT literature, although some authors— especially those focusing on the Rasch model—employ different nomenclature); operationally, it is defined as the score on θ that is associated with a 50% likelihood of a correct /endorsed item response. In the hypothetical item presented in Figure 1, the *b* parameter would equal 1.0 (i.e., the point at which the ICC intersects the horizontal reference line at PCR = 0.50).

It should be noted that the difficulty parameter (*b*) and θ lie on the same scale due to the fact that the former is defined directly in terms of the latter. The fact that the *b* parameter (which defines a characteristic of a test item) lies on the same scale as θ (which defines a characteristic of a person being assessed) represents a very important characteristic of IRT models; that is, they locate these person and item parameters on a common scale. In contrast, CTT-based item parameters (e.g., proportions of correct responses, item discrimination correlations) lie on a very different scale than that used to estimate each respondent's score on the trait in question.

By implication, in the 1-parameter model, all items in a test exhibit ICCs having the same shape; the only characteristic that distinguishes one item's ICC from another is the left-right location of the ICC on the horizontal axis (θ), that is, its "difficulty." Figure 2 presents ICCs for three hypothetical items with *b* parameters of –1.0, 0, and 1.0. As an inspection of these three ICCs illustrates, the form of the functional relationship between θ and the observed response (i.e., the shape of the ICC) is constant across items; all that differs is the level of θ that is associated with a given observed probability of a correct / keyed response.

For example, Item A (solid line) is the least difficult of the three items; even in a group of individuals having a relatively low score on θ (i.e., –1.0 *z* units, or approximately the 16th percentile in a normal distribution)—that is, the HSP located at –1.0—fully 50% of these individuals would be expected to provide correct responses to this item. In contrast, only about 17% of indi-

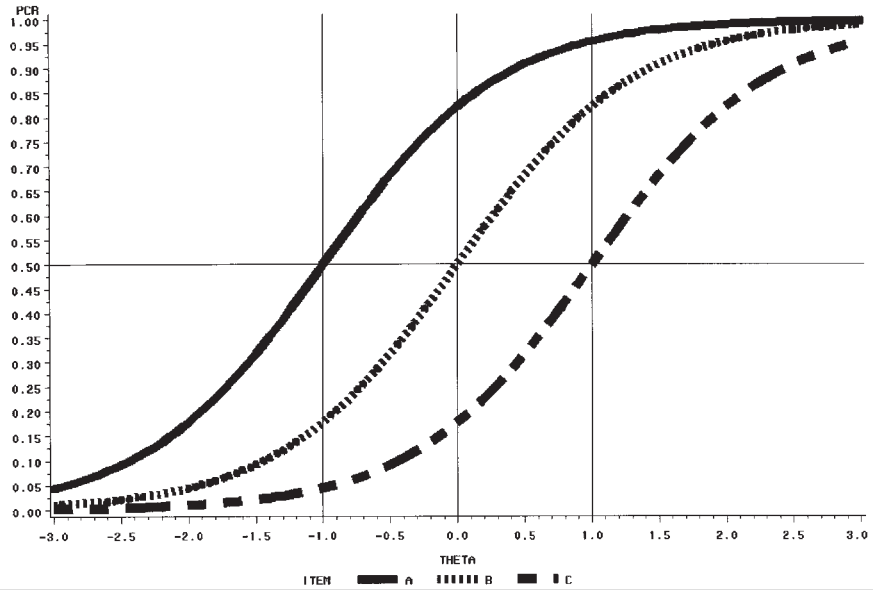

Figure 2. Item characteristic curves for three hypothetical test items.

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the observed probability of a correct response, in the case of a right-wrong item, or the probability of a response in the keyed direction. These items have identical parameters of $a = 0.9$ and $c = 0$; they differ with respect to b, with values of $\theta = -1.0$, 0.0, and 1.0 for items A, B, and C, respectively.

viduals in the –1.0 HSP would be expected to be able to get Item B correct, and only approximately 5% of them would be expected to provide correct responses to Item C.

2-parameter logistic model. The main potential drawback to the 1-parameter IRT model is its assumption that all items in the test share identically shaped ICCs; although this might be attainable in an item pool that was very carefully selected from a much larger initial pool of items, it would be quite unusual in many applied assessment situations. In response, the 2-parameter model adds a parameter—termed *discrimination*, or *a*—that allows the ICCs for different items to exhibit different slopes. The discrimination parameter allows us to model the fact that some items have stronger (or weaker) relations than others to the underlying construct being assessed (θ); larger values of *a* denote stronger relations (i.e., in somewhat the same way that in a factor-analytic context, items that demonstrate larger loadings on a factor are seen to be more strongly related to that factor than are items exhibiting smaller loadings). The *a* parameter is very important in IRT due to the fact that it directly determines the amount of information provided by an item: Items with higher *a* parameters provide more information regarding θ, all other factors being equal. The ICCs for three hypothetical items having identical difficulty, but different discrimination, are presented in Figure 3.

As can be seen in Figure 3, the IRT model allows great freedom with regard to the way in which θ and the item parameters can combine to produce different observed item-response patterns; in the case of these three items, very different results (i.e., expected item-endorsement patterns) would be expected depending on which HSP was being considered. For example, in the HSP containing individuals who score at θ = –1.0, we would expect that Item C (the most highly discriminating item) would be answered correctly/endorsed at the lowest rate (approximately 7%) in this group, with Item B (moderate discrimination) correct /endorsed at a higher rate (15%), and the least-discriminating Item A being correct /endorsed at the highest rate (approximately 30%). In contrast, in the HSP at θ = 0.0, we would expect all three items to be correct /endorsed at the same rate (50%), and in the HSP at θ = 1.0, we would expect the pattern seen in the HSP at θ = –1.0 to be reversed (i.e., the most discriminating item would be correct/endorsed at the highest rate, and the least discriminating item at the lowest rate). Clearly, if the situation depicted in Figure 3 were found in practice, highly misleading conclusions could be reached if the overly simplistic model being fit by the 1-parameter, or Rasch, approach were applied to such data.

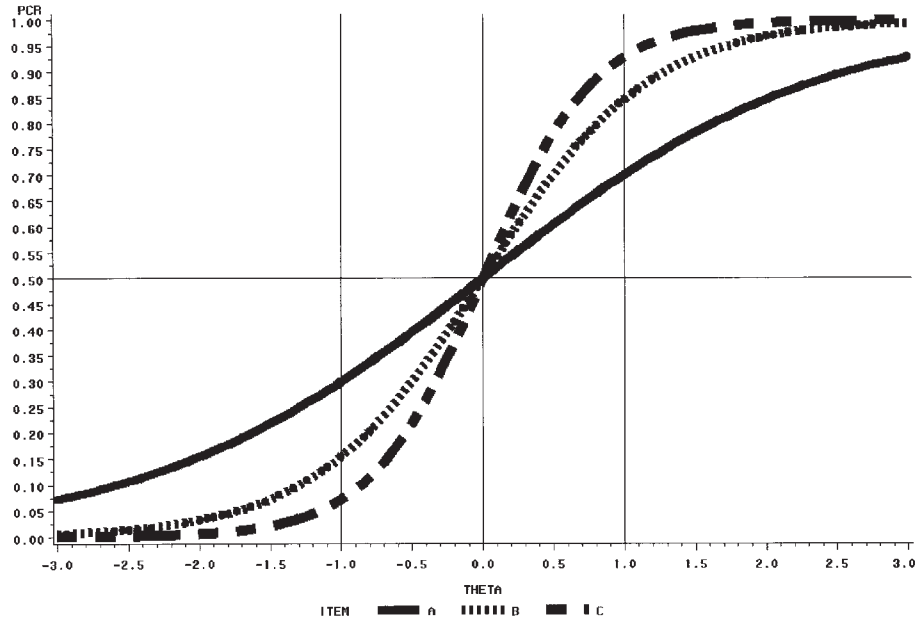

Figure 3. Item characteristic curves for three hypothetical test items.

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the observed probability of a correct response, in the case of a right-wrong item, or the probability of a response in the keyed direction. These items have identical parameters of $b = 0.0$ and $c = 0$; they differ with respect to a, with values of 0.5, 1.0, and 1.5 for items A, B, and C, respectively.

3-parameter logistic model. Although the 2-parameter model addresses one of the most serious criticisms of the Rasch model (i.e., the assumption that all test items are identical with respect to their discriminating power), it does not address another potentially important factor that may differ across items: namely, the lower asymptote of the ICC (i.e., the expected proportion of correct/keyed responses that would be expected from individuals who have very low θ scores). The 3-parameter model adds one more parameter (*c*) to the 2-parameter model to reflect the fact that the lower asymptote of the ICC may need to adopt nonzero values for their effective minimum values (i.e., in both the 1- and 2-parameter models, the lower asymptote of the ICC is fixed to zero).

Initially, when IRT models were developed in the context of right-wrong tests, the main reason for postulating the need for a nonzero lower asymptote was the fact that in tests composed of multiple-choice items, individuals who did not know the correct answer could be expected to guess, and they would occasionally guess the correct response. Thus, even HSPs composed of examinees with extremely low θ scores might well be expected to produce decidedly nonzero rates of "correct" responses to difficult test items due to guessing. Later, when IRT models began to be applied to tests composed of items that were not rated in a right-wrong fashion (e.g., Harvey & Murry, 1994, used the 3-parameter IRT model to analyze the keyed responses to the MBTI), although guessing was not a particular concern, it was still possible that some items' ICCs would not exhibit zero lower asymptotes due to social desirability or the relatively extreme nature of some items.

Figure 4 presents three hypothetical ICCs for items with identical *a* and *b* parameters (1.0 and 0.0, respectively), differing only in their *c* parameters (0.0, 0.25, and 0.5), representing the situation in which effectively no guessing or social desirability was present or possible (Item A, which, for example, if this were a right-wrong item, might occur in a non-multiple-choice, open-ended response item in which the chance of guessing the correct response was very small); a moderate level of guessing was expected (Item B, which might correspond to a 4-alternative multiple-choice item); and guessing was very easy (Item C, which might correspond to a 2-alternative multiple-choice item, such as would be seen in true-false, endorsed–not endorsed, or checklist-format items).

The ICCs in Figure 4 illustrate the unavoidable effect of increasing the *c* parameter in the 3-parameter IRT model: namely, reducing the effective discriminating power of an item (and thereby reducing the level of information it provides while simultaneously decreasing its effective level of difficulty in a CTT-based sense, wherein difficulty is defined in terms of the overall percentage of correct, or keyed, item responses). In the context of a right-wrong item, this concept is relatively easy to understand anecdotally; that is, the

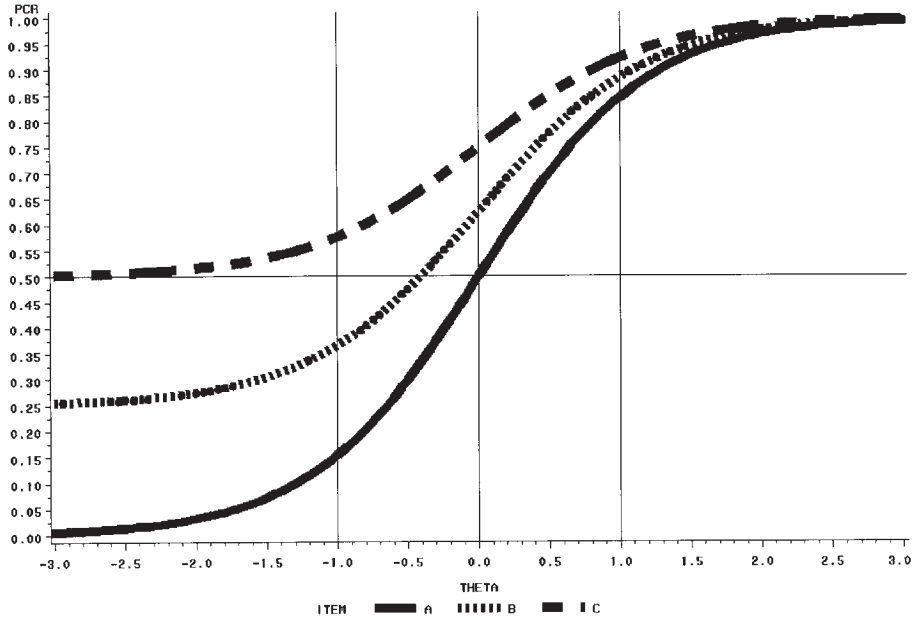

Figure 4. Item characteristic curves for three hypothetical test items.

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the observed probability of a correct response. These items have identical parameters of $b = 0.0$ and $a = 1.0$; they differ with respect to c, with values of 0.0, 0.25, and 0.5 for items A, B, and C, respectively.

easier it becomes to guess the correct response to an item, the less informative that item becomes with respect to providing us with information that is useful in estimating the θ score for the person completing the test. Conversely, the more difficult it becomes to guess the correct answer, the more diagnostic a correct (or incorrect) answer becomes.

Differences Between IRT and CTT

Most of the tests and inventories used by counseling psychologists have been developed using CTT; IRT derives from what is called "modern test theory" and is one of the methodologies that have resulted in what Embretson (1996) calls "the new rules of measurement." A number of important differences exist between CTT-based and IRT-based approaches to both test development and evaluation as well as the process of scoring the response profiles of individual examinees; these differences are summarized below.

Item-level focus. Although tests have always been composed of multiple items, IRT takes a much more item-level focus than CTT, which tends to focus more on test-level indices of performance (e.g., the overall reliability coefficient, or standard error, of a scale). In particular, the focus on estimating an ICC for each item provides an integrative, holistic view of the performance of each item that is not readily available when using CTT-based methods to develop or examine a test. That is, although CTT can quantify the total-sample difficulty (e.g., as a *p* value) or discrimination (e.g., as an item-total biserial correlation) for an item, it lacks an effective means for simultaneously combining and presenting this information (including the role of guessing or other factors that might lead to a nonzero lower asymptote) in an easily used format.

Continuous view of information and standard error. The concept of information in IRT is roughly analogous to the concept of reliability in CTT in the sense that higher levels denote better measurement precision (or conversely, freedom from undesirable errors of measurement). In IRT, higher levels of information are produced when items have higher discrimination (*a*) parameters and smaller lower asymptote (*c*) parameters. Similarly, the standard error associated with a particular θ score estimate is inversely related to information in IRT, much as the standard error of measurement (*SEM*) is inversely related to the reliability of a test in CTT.

The critical difference is that in IRT we need not assume that the test is equally precise across the full range of possible test scores, as is effectively the case when CTT-based methods are used. Whereas in CTT a single number (e.g., the internal-consistency reliability coefficient, or the *SEM*

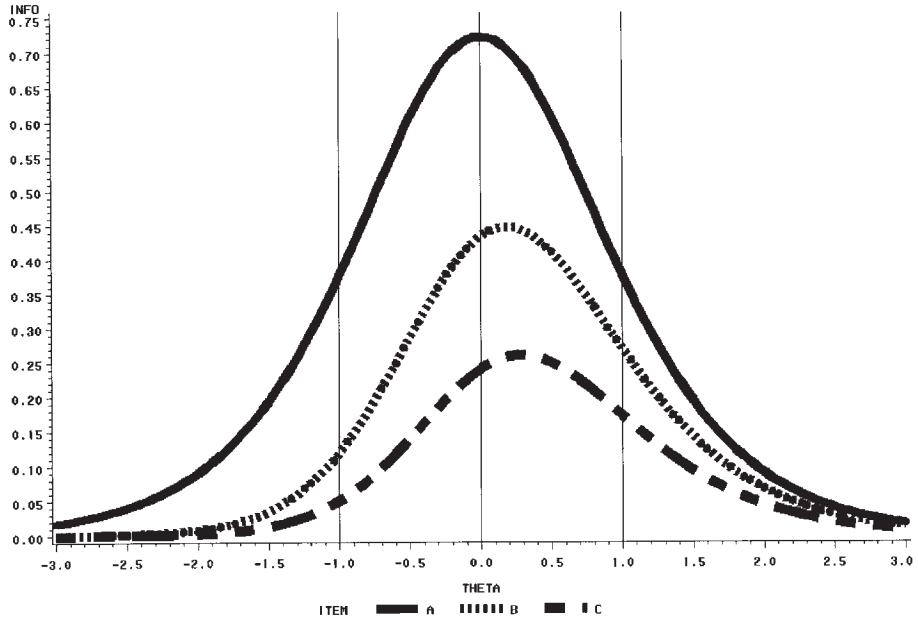

Figure 5. Item information functions for the three hypothetical test items presented in Figure 4.

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the amount of information provided by each item.

based on that reliability) would be used to quantify the measurement precision of a test, a continuous function is required in IRT to convey comparable data, given that in the IRT approach, a test need not be assumed to possess a constant degree of measurement precision across the entire possible range of scores. Figure 5 presents the item information functions (IIFs) for the three hypothetical items presented in Figure 4.

The IIFs presented in Figure 5 illustrate two important aspects regarding item-level information when using the 3-parameter IRT model: (a) The height of the maximum point in the information function is directly reduced by the *c* parameter, and (b) the location of the point of maximum information is shifted rightward from the value that would be expected in the 1- and 2-parameter models (i.e., at $\theta = b$) in an amount that is proportional to the size of *c*.

Test development: item selection, scoring. The IRT-based approach to test development has the advantage of allowing the test developer to easily determine the effect of adding, or deleting, a given test item or set of test items by examining the test information function (TIF) and/or test standard error (TSE) function for an item pool. The TIF is computed simply as the sum of the IIFs for the items in the pool being examined; by examining the change in the shape of the TIF or TSE functions after adding or deleting items, and comparing this to the desired performance curve, tests can be tailored closely to desired specifications. Under CTT, only much less sensitive measures (e.g., the global coefficient alpha or *SEM* for a given test) are available.

With respect to test scoring, IRT-based methods—especially those based on the 2- or 3-parameter models—offer considerable advantages over the "number-right" scoring methods typically used in CTT-based tests. Specifically, when estimating an examinee's score using IRT, we can simultaneously consider the following sources of information: (a) which items were answered correctly or incorrectly (or in the keyed vs. nonkeyed direction) and (b) for each of those items, the difficulty, discrimination, and nonzero lower asymptote parameters of the item. This offers the potential to produce better estimates of the θ scores, to produce quantitative estimates of the "quality" or likelihood of any given observed response profile (termed *appropriateness indices*) (e.g., Drasgow, Levine, & McLaughlin, 1987), and to assess the degree to which the given IRT model provides a good fit to the pattern of responses produced by the individual in question.

Differential item functioning. One of the important issues faced by counseling psychologists is that of responding to the diversity of clients. In particular, it is important that the tests used by counseling psychologists be free

of systematic demographic subgroup bias. IRT techniques provide a powerful means of testing items for bias, using what is known as *differential item functioning* (DIF), as well as assessing the cumulative effect of any item-level bias on the test's total score (which in many situations is the most important bottom-line issue) (e.g., Drasgow, 1987).

In contrast, CTT-based methods of assessing bias are fundamentally limited, especially approaches that base their assessments of bias on the presence of group mean differences in total test scores across demographic groups or on differential item-passing/endorsement rates between subgroups (e.g., Drasgow, 1987). In essence, such methods cannot distinguish between the situation in which (a) the subgroups have different means, and the test is biased, versus (b) the means differ, but the test is not biased (i.e., one group truly has a higher average on the test).

Computer-adaptive testing. One of the potentially most important differences between CTT-based and IRT-based testing concerns the issue of administrative efficiency (i.e., reducing testing time) and item banking (i.e., developing calibrated item pools from which subsets of items can be selected for each individual tested). Whereas CTT-based indices of test functioning—and especially, scoring—are fundamentally based on the assumption that the entire item pool is going to be administered to each examinee, IRT-based methods can easily deal with the situation in which different examinees are presented with entirely different listings of items or different numbers of items. This is due to the fact that the scoring methods used in IRT to estimate each examinee's θ score can produce estimates that lie on a common θ score metric even if there is little—or no—overlap between examinees in terms of the test items that are administered; in contrast, the number-right scoring methods typically used in CTT-based approaches are highly dependent on having the same list of items be presented to each examinee.

A growing number of implementations of IRT-based tests (e.g., Sands et al., 1997) have demonstrated that reductions in testing time of up to one half can be achieved by using CAT methods to tailor the administration of test items to the estimated level of θ for each examinee without compromising the measurement precision or security of the test. This is a tremendously important feature for counseling psychologists, as it allows them either to achieve dramatic reductions in overall testing time, to administer a larger number of different tests in the same amount of time as was required using nonadaptive administration methods, and to tailor the selection of test items so as to produce a test that produces its highest degree of measurement precision in a specified range of the θ scale.

APPLICATION: THE USE OF IRT TO REVISE THE MYERS-BRIGGS TYPE INDICATOR

To illustrate the relevance of IRT to counseling psychologists, we consider the recent revision of the MBTI (Myers et al., 1998). This example represents a real-life application of how IRT was used to revise a widely used instrument. There are a number of reasons why the MBTI revision provides a good example of IRT techniques. First, the MBTI is generally well known by counseling psychologists (Graff, Larrimore, Whitehead, & Hopson, 1991). Most practitioners or researchers in the field, even if they do not use the instrument themselves, are at least likely to be familiar with its concepts and the associated measurement issues. Second, this example shows how IRT can be used to develop or revise a personality instrument; although frequently applied to right-wrong, aptitude, and achievement tests, IRT methods have not been as widely used with other instruments, even those that lend themselves easily to dichotomous item-level keying (e.g., the MBTI). Third, the use of IRT with the MBTI demonstrates the broad applicability of IRT to different kinds of measurement problems because the MBTI is a theory-driven instrument designed to measure types rather than traits.

The MBTI is composed of four bipolar preference scales that attempt to measure Jung's theory of psychological types. Each scale is composed of a set of forced-choice items, with the four scales being Extraversion-Introversion (E-I), Sensing-Intuition (S-N), Thinking-Feeling (T-F), and Judging-Perceiving (J-P). There were two primary measurement problems faced by those working on the MBTI revision. The first involved the issue of how to select items for the instrument; the second was the issue of how to score those items to arrive at a four-letter categorical type.

In all cases, given the lack of a right-wrong answer, item responses were keyed, with a 1 response being given if the item was endorsed in the I, N, F, or P direction and with a 0 response for answers in the E, S, T, or J direction. The choice of a keyed direction is completely arbitrary, having the effect of simply setting the direction of the θ scale for each of the four MBTI scales.

Item Selection

The method previously used in the MBTI to select items and provide a classification on each scale employed what were called *prediction ratios*. A prediction ration was computed for each response to each MBTI item by dividing the percentage of people holding the target preference who answered an item in the keyed direction (e.g., a person with a preference for Thinking who chose the response keyed to Thinking) by the percentage of everyone answering that item in the keyed direction. Based on previous

research, Myers selected items for inclusion on the MBTI if the prediction ratio for at least one of the responses was greater than .62 (e.g., Myers & McCaulley, 1985). There were other criteria as well, including rejecting any item for which greater than 50% of the people with the nontargeted preference responded in the keyed direction, item-to-scale correlations, and linguistic and theoretical criteria. These latter, however, are not relevant to this example.

The first step in the use of IRT to select items for the revised MBTI involved the calculation of the item parameters for the 3-parameter IRT model described above. IRT parameters can obviously be used to select items with different properties and tailor item selection to a given purpose; the important question in the context of the MBTI revision was, What properties of MBTI items would be consonant with the MBTI theory? This theory posits that a person will demonstrate a preference for one pole or the other of each of the four bipolar scales and that this preference represents a qualitative difference, not a quantitative one.

It was therefore desirable to have items that would help sort people into the correct qualitative preference categories, which in IRT terms is essentially a problem of choosing items whose maximum amount of information, or discrimination, occurred around the midpoint of each scale (the midpoint of the θ scale approximating the cutoff point that would assign examinees into the categorical types). Given the theory on which the MBTI is based, "good" items would be those that demonstrated ICCs like the one seen in Figure 6, which represents the empirical ICC produced for Item 3 in Form M, a high-performance item from the E-I scale (i.e., "quiet and reserved," the I keyed response, vs. "good mixer," the E nonkeyed response).

The ICC depicted in Figure 6 demonstrates that for this item the measurement model underlying IRT fits the MBTI data very well. That is, the ICC is clearly nonlinear, and there is relatively little "scatter" around the nonlinear regression line fit through the item-endorsement percentages computed in each of the HSPs. People whose preference is for Extraversion in general have little probability of selecting the response keyed to Introversion ("quiet and reserved"), and vice versa. For example, considering those HSPs that endorsed fewer than 20% of the total number of E-I items in the keyed (I) direction— that is, very clear Extraverts in MBTI terminology—we see virtually zero rates of endorsement of the Introvert alternative to this item. Likewise, among clear Introverts—those endorsing 80% or more of the items in the I direction— we observe endorsement rates for the keyed response in the 100% range. In the transition region, the ICC exhibits a sharp slope, with a very good approximation being provided by the nonlinear regression line.

According to Myers's type theory, it is the preference itself, and not the score or strength of the preference, that primarily determines the likelihood

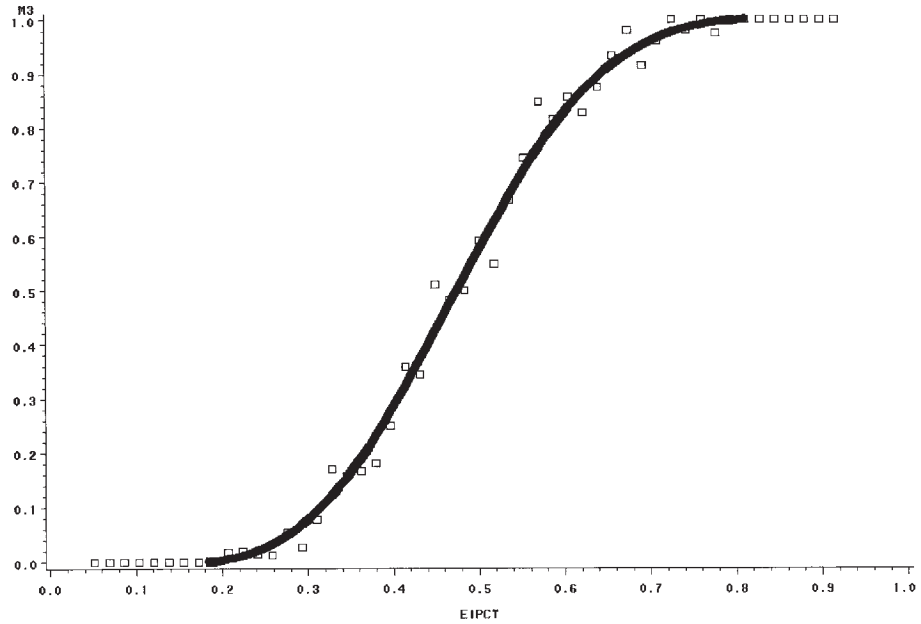

Figure 6. Empirically derived item characteristic curve for Myers-Briggs Type Indicator Item 3 in Form M ("good mixer " vs. "quiet and reserved").
NOTE: The horizontal axis represents the preference on the Extraversion-Introversion (E-I) dimension, with the vertical axis representing the percentage of respondents in each subgroup (formed using the total number of endorsed items) that endorsed the item in the keyed (i.e., I) direction. The squares represent the percentage of raters in each homogeneous subpopulation endorsing this item in the keyed (I) direction.

that a person will respond to the item in the keyed direction. From that perspective, the item in Figure 6 exhibits precisely the kind of ICC that would be desired. That is, at the high (I) and low (E) ends of the θ scale, respondents are near unanimous with respect to their patterns of endorsing (or not endorsing) the item in the predicted direction; only in the middle ranges of the θ scale do we find intermediate levels of endorsement, and this transition zone is relatively narrow.

Without the use of IRT when constructing or revising this scale, we might run the danger of choosing items like the one shown in Figure 7; this item was included in the tryout form of the revised MBTI (Form X), but it was not retained for use in the final Form M. The primary reason this item was not retained can be seen in Figures 7 and 8: Figure 7 shows that the ICC for this item exhibits a very different shape than the ICC for the high-performance item depicted in Figure 6; in particular, it has a much more shallow slope (hence, less discriminating power). Figure 8 presents the IIFs for these two items, showing quite graphically that the item in Figure 6 produces far more information regarding θ—especially in the middle region of the θ scale, where the type cutoff point would be located—than does the item in Figure 7. Accordingly, good items from the perspective of revising the MBTI using IRT methods are those that produce the most information and that produce their maximum information in the middle region of the θ scale (i.e., so as to maximize test precision near the type cutoff point).

Note that a "good" item for the purposes of the MBTI does not necessarily have to assume the shape of a step function (i.e., an ICC in which all people holding the nonkeyed preference respond in the nonkeyed direction and all people holding the keyed preference respond in the keyed direction, with essentially zero transition from one to the other). Although there may be some items that begin to approach this goal (such as the one shown in Figure 6), it is not required that all items assume this ideal form because neither Jung nor Myers believed that everyone was a "type." The fact that even the best of the MBTI items (i.e., the ones with the highest discrimination parameters) do not precisely conform to the ideal of a step-function ICC demonstrates the theoretical proposition that for various reasons, including developmental or situational factors, a given person at a given time will have some chance of responding to an item in a direction opposite that of his or her preference.

It is important to note that the criteria outlined above for MBTI are not necessarily those that would be adopted if one were interested in discriminating among people at many different points along a scale or across the entire possible range of θ values. For example, a psychologist designing an aptitude or achievement test would probably want a set of items whose maximum information occurred at evenly spaced points along the continuum of latent

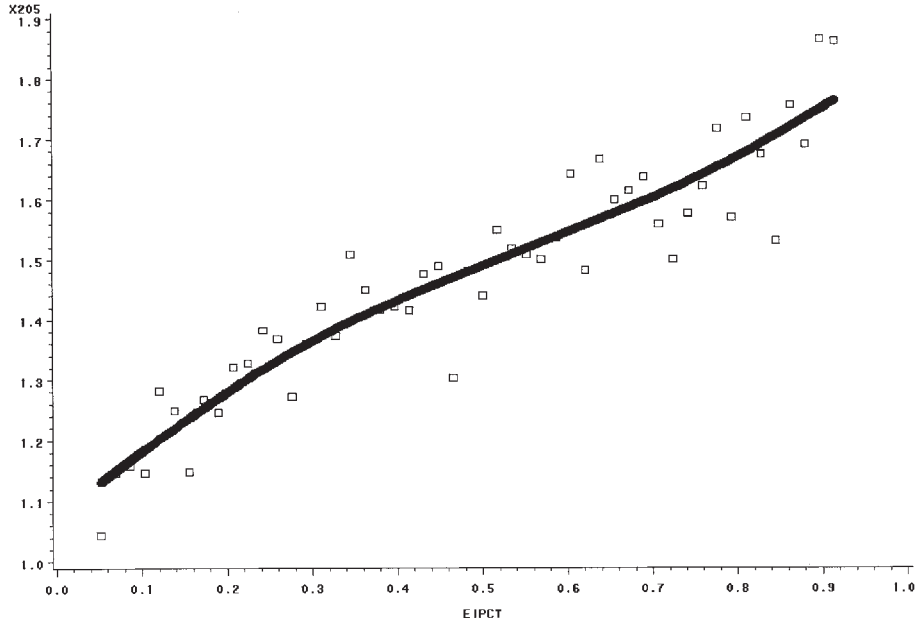

Figure 7. Empirically derived item characteristic curve for Myers-Briggs Type Indicator Item 205 in the experimental Form X (in business situations, "sticking to business" vs. "adding extra socialization").

NOTE: The horizontal axis represents the preference on the Extraversion-Introversion (E-I) dimension, with the vertical axis representing the percentage of respondents in each subgroup (formed using the total number of endorsed items) that endorsed the item in the keyed (i.e., I, or "sticking to business") direction. The squares represent the percentage of raters in each homogeneous subpopulation endorsing this item in the keyed (I) direction.

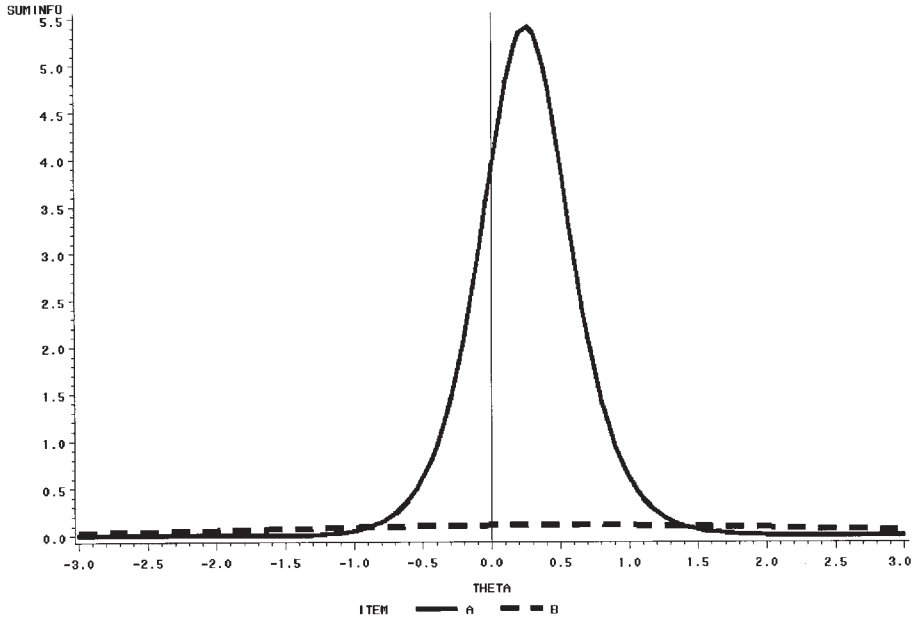

Figure 8. Item information functions for the two items depicted in Figure 6 (Item A) and Figure 7 (Item B).

NOTE: The horizontal axis represents θ—the preference on the Extraversion-Introversion (E-I) dimension—with the vertical axis representing the amount of information provided by the item at each point along the θ scale.

achievement levels, given that the test's goal would be to produce the most precise estimate of θ for the largest range of possible examinees.

For the MBTI revision, having established the general criteria of *S*-shaped curves whose maximum slope occurred somewhere close to the midpoint of the scale, the next step was to estimate the three IRT parameters for each item, using the initial item pool contained in the tryout version of the instrument. The sample used to estimate the three parameters was a national random sample of adults (N = 3,000) (Myers et al., 1998). Items whose parameters produced ICCs of the general shape as those in Figure 6 were selected for further research and were subjected to additional empirical and theoretical criteria; those producing ICCs more like Figure 7 were discarded.

A number of items were also discarded as a result of DIF analyses using subgroups for gender and for three age-based groupings. For example, consider the MBTI item, "Do you usually (A) show your feelings freely, or (B) keep your feelings to yourself?" Based on the traditional prediction-ratio method, this item was used on the E-I scale of an earlier version of the MBTI. Although the response scoring weights indicated that it was not a highly discriminating item, its power was sufficiently good to cause it to be retained; additionally, it did not employ differential scoring weights for men and women (only selected items on the T-F scale used differential scoring).

However, DIF analysis (see Figure 9) revealed that the item responses for this item were significantly different for men and women, and consequently, this item was dropped from the revised form. As the subgroup ICCs presented in Figure 9 illustrate, even when males and females shared identical standing with respect to the E-I preference (θ), males were consistently more likely than females to endorse the response keyed in the Introvert direction. For example, consider the HSP of respondents holding scores of –1.0 θ units (i.e., relatively clear Extraverts); approximately 40% of the males in this subgroup endorsed the Introvert response to this item, whereas only about 30% of females holding the identical θ score endorsed the Introvert response. This pattern of males being approximately 10% more likely to endorse the keyed (Introvert) response to this item is consistent across most of the θ scale, indicating that some sort of systematic difference between males and females other than their scores on the E-I scale was affecting their responses to this item. In this case, it is tempting to speculate that this difference is a function of the emphasis on feelings in the item. That is, in addition to measuring the E-I preference, this item may also effectively function as an indicator of the T-F preference. In view of the fact that males tend to score lower, on average, than females on the T-F preference (e.g., Myers & McCaulley, 1985), it is not surprising to find that males would be more likely to endorse the "nonfeeling" alternative (i.e., the one keyed toward Introversion), all other things (including their scores on the E-I preference) being equal.

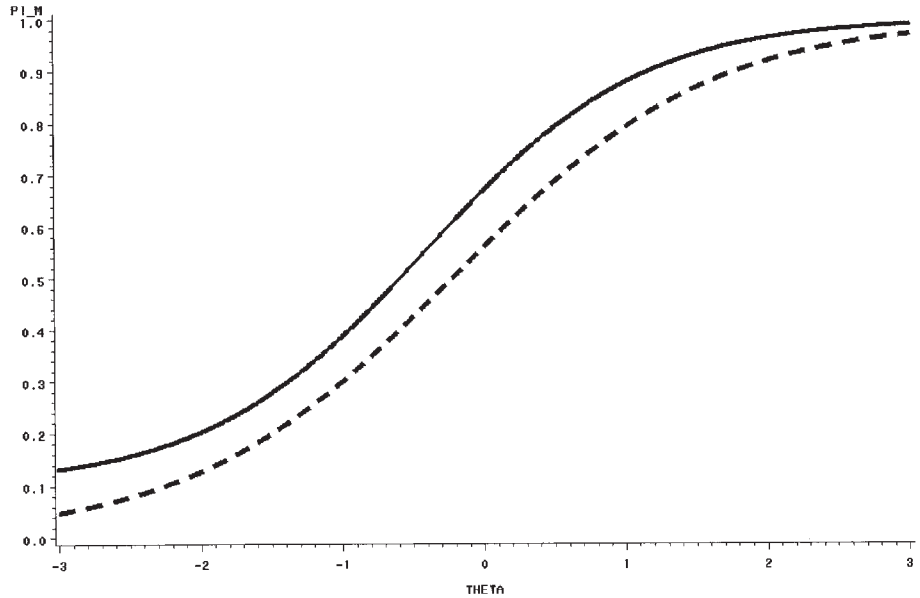

Figure 9. Subgroup item characteristic curves (ICCs) for males (solid line) versus females (dashed line) for Extraversion-Introversion (E-I) Item 15 from Myers-Briggs Type Indicator Form F/G (keyed, I, response = "keep your feelings to yourself," vs. nonkeyed, E, response = "show your feelings freely")

NOTE: The horizontal axis represents θ (the underlying construct to be measured), with the vertical axis representing the observed probability of a response in the keyed direction. These ICCs reveal differential item functioning such that males are consistently more likely to endorse the keyed (I) response than are females who share the same underlying E-I θ score.

In summary, the use of IRT DIF analysis during the MBTI revision led to the following conclusions: (a) Items in the original item pool that demonstrated significant gender differences were not confined to the T-F scale, (b) some of the items with separate weights on the T-F scale on the previous version did not show significant DIF, (c) there was a small number of items that showed age-related DIF, and (d) once the gender- and age-related items that showed significant DIF were dropped from the scales, it was possible to demonstrate that there was no DIF at the scale level, which is the level at which classifications decisions are made. Although ultimately affecting only a small portion of respondents, all of these findings represented an advance in the understanding of MBTI item functioning due to the use of IRT DIF methods.

Application of the IRT analyses, including the DIF methods, along with additional theoretical and empirical criteria, resulted in the selection of 93 items to comprise Form M of the MBTI. Forty-three items included on the previous version of the MBTI were dropped and replaced with revised items whose ICCs were more like the *S*-shaped curves of Figure 6.

Scoring

IRT is used to score MBTI items using a weighted maximum-likelihood approach (e.g., Bock & Mislevy, 1982) that is sensitive not to simply the number of items endorsed in the keyed direction but to the simultaneous considering of the direction of the item response and the *a*, *b*, and *c* parameters for each item. The basic maximum-likelihood scoring method produces a likelihood function for each examinee that ranges across the entire possible scale of θ and shows how likely a given θ score would be given the observed pattern of item responses produced by that person and considering our knowledge of the *a*, *b*, and *c* parameters for the items that were administered. By examining this likelihood function, we can identify the most likely value of θ that would be consistent with the observed item responses and use this value as our estimate of θ for that person. Figure 10 presents likelihood functions for three profiles of responses to the E-I MBTI items.

In Figure 10, the solid line represents the likelihood function for an individual who holds a preference in the nonkeyed (E) direction on the E-I scale, as seen by the fact that the peak of the likelihood function occurs to the left of the middle point of the scale; likewise, the function for the dotted line depicts a person who holds the keyed (I) preference, given that the location of the maximum occurs to the right of the type cutoff point. The height of the likelihood function, at its maximum, also provides important information in IRT scoring: the higher the maximum, the more likely that the θ score esti-

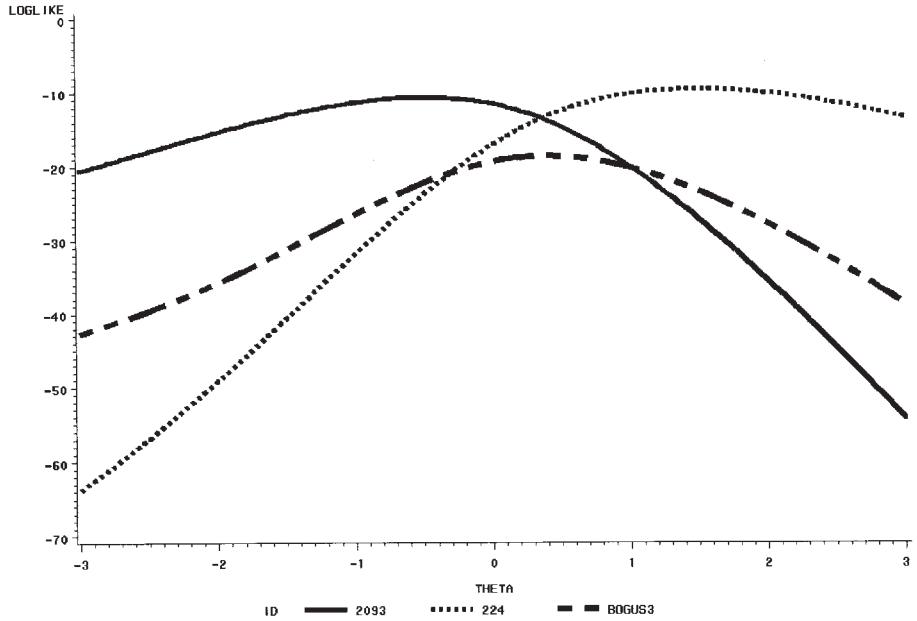

Figure 10. Log-likelihood functions for three Myers-Briggs Type Indicator response profiles to the Extraversion-Introversion (E-I) scale.
NOTE: IDs 2093 and 224 are based on actual item ratings, whereas the third profile was simulated using random item responses. The horizontal axis represents θ (the preference on the E-I dimension), with the vertical axis representing likelihood that the observed pattern of item endorsements would have been seen from a person holding the given θ score (higher values are more likely).

mate is indeed accurate and consistent with the underlying IRT model being assumed.

The third profile in Figure 10 (heavy dashed line) was obtained using random responses to the E-I items; comparison of this likelihood function against the two produced using real data reveals that the random-data profile does not exhibit a maximum value that is clearly toward one pole or the other, and of greater importance, the height of the likelihood function for the random-data profile is significantly lower than the height for either of the two real-data profiles. This height of the likelihood function can be used (e.g., Drasgow, Levine, & McLaughlin, 1987) to produce appropriateness indices that quantify the degree of consistency or plausibility of each observed item-response profile; such indices are of potentially significant practical importance (e.g., in detecting invalid, questionable, faked, or erroneous profiles).

SUMMARY

Benefits of the IRT approach include the fact that it provides a much more detailed view of item-level and test-level functioning (e.g., with respect to information and standard errors); it can be adapted to many different kinds of tests; the score estimation process is more precise, allowing simultaneous consideration of both the number of right/endorsed items as well as the properties (e.g., discrimination, difficulty) of each item, when estimating each person's score on the construct being assessed; the degree to which the IRT model fits consistently across demographic subgroups of respondents (e.g., males vs. females) can be assessed in order to document the lack of subgroup bias in a test (e.g., Gratias & Harvey, 1998; Harvey, 1997); fit statistics that quantify the plausibility of each observed item-response profile can be calculated and used to target responses for closer examination; and CAT methods can be used to dramatically reduce test administration time if the test can be administered in a computer-based format (e.g., Laatsch & Choca, 1994; Sands et al., 1997; Waller & Reise, 1989).

Of course, when we use IRT to develop and score tests, it is important to remember that we are indeed fitting a specific mathematical model to our empirical data and that this model incorporates certain assumptions; as with any other multivariate data analytic method, certain caveats need to be kept in mind. First, there is no guarantee that the model underlying the given IRT approach (e.g., 1-parameter vs. 2- parameter vs. 3-parameter logistic) will indeed provide an adequate degree of fit to the data; obviously, to the extent that the true relation between item responses and the underlying construct or constructs of interest do not follow the form that is assumed by the IRT model, difficulties in interpretation will arise. For example, consider the Rasch, or

1-parameter, model: Despite its continued popularity in some quarters (particularly educational measurement) (e.g., Wright, 1977), the Rasch model is based on very restrictive assumptions regarding the nature of the data being analyzed (e.g., that all items have identical discriminating power and that absolutely no factors that would cause nonzero lower asymptotes of the ICCs, such as guessing, are operative). Although there may be some situations in which such restrictive assumptions provide an accurate representation of the processes producing test item responses, the number of such situations in applied assessment settings are probably limited. As when fitting any model to data, researchers should always attempt to empirically evaluate the degree to which the IRT model being used actually provides an acceptable fit to the data being analyzed (e.g., see Hambleton et al., 1991). Indeed, the fact that IRT models offer a variety of powerful methods for assessing the degree of model fit—at both the person and item level of analysis—can be seen as a powerful advantage over traditional CTT-based methods.

Another potential issue—although it is certainly not unique to tests that are scored using IRT methods—centers on IRT's assumption that the scales being measured are unidimensional in nature. IRT methods may not be applicable in situations in which this assumption is significantly violated, such as when external-criterion-based methods of item analysis are used to select and score the test items (such as occurs with some personality tests derived using empirical keying) (e.g., Waller & Reise, 1989). On the other hand, research has suggested that IRT analyses are reasonably robust with respect to violations of the unidimensionality assumption, leading Drasgow and Hulin (1990) to suggest that such violations be treated with the same degree of concern as violations of assumptions in other standard analytic techniques such as regression or ANOVA.

The MBTI again provides a practical example of the application of this assumption in a real case. In the MBTI revision process described above, IRT analyses for item selection and scoring were developed separately for each of the four MBTI preference scales. In these analyses, each scale was assumed to be unidimensional despite the fact that research has shown that each of the four MBTI preference scales can be partitioned into additional subfactors (Johnson & Saunders, 1990). However, research has also shown (Harvey, Murry, & Stamoulis, 1995) that the four MBTI item pools are each dominated by a single underlying construct; past research (e.g., Drasgow & Parsons, 1983; Hulin et al., 1983) has suggested that this is the important factor, not that one be able to demonstrate that a one-factor model can account for 100% of the variance in a given item pool. As analyses such as the empirical ICCs presented in Figure 6 demonstrated, the 3-parameter IRT model was indeed able to provide a good fit to the MBTI data in each of its four scales.

A final limitation on the use of IRT concerns the need for large samples and relatively large numbers of items in each scale. In the early days of IRT, very large samples were needed for practitioners to have a reasonable chance of obtaining accurate and stable estimates of the unknown item parameters of the IRT model. Fortunately, improved statistical algorithms (e.g., Bock & Aitkin, 1981) have proven effective in reducing this as a concern, to the point that Drasgow and Hulin (1990) suggested that the sample size requirements for IRT are comparable to that of factor analysis and similar multivariate methods. Continued refinements in available software programs to implement IRT analyses with respect to performance and ease of use are likewise occurring (e.g., Gierl & Ackerman, 1996).

Taken as a whole, then, IRT offers a tremendous degree of promise as a powerful and flexible method for test development, scoring, and evaluation; as many authors have noted (e.g., Embretson, 1996; Hambleton et al., 1991; Sands et al., 1997), IRT methods represent a vast improvement over approaches based on CTT. They are certainly not free from potential concerns or limitations, although no data-analytic method ever will be.

REFERENCES

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-445.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Brown, R. D., & Harvey, R. J. (1998, April). *Computer-adaptive testing and test-retest reliability in a "Big-Five" personality inventory.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas.

Drasgow, F. (1987). A study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology*, *72*, 19-29.

Drasgow, F., & Hulin, C. L. (1990). Item response theory. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial & organizational psychology* (2nd. ed., Vol. 1, pp. 577-636). Palo Alto, CA: Consulting Psychologists Press.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189-199.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349.

Gierl, M. J., & Ackerman, T. (1996). Software review: XCALIBRE—Marginal maximum-likelihood estimation program, Windows version 1.10. *Applied Psychological Measurement*, *20*, 303-307.

Graff, R. W., Larrimore, M., Whitehead, G. I., & Hopson, N. W. (1991). *Career counseling practices: A survey of college/university counseling centers*. Poster presented at the annual meeting of the American Psychological Association, San Francisco.

Gratias, M., & Harvey, R. J. (1998, April). *Gender and ethnicity-based differential item functioning in the Myers-Briggs Type Indicator.* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Harvey, R. J. (1997, April). Computer adaptive testing, differential item functioning, faking, and the MBTI. In R. J. Harvey (Chair), *Using item response theory to address assessment challenges in I/O*. Symposium conducted at the Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis.

Harvey, R. J., & Murry, W. D. (1994). Scoring the Myers-Briggs Type Indicator: Empirical comparison of preference score versus latent-trait methods. *Journal of Personality Assessment*, *62*, 116-129.

Harvey, R. J., Murry, W. D., & Stamoulis, D. (1995). Unresolved issues in the dimensionality of the Myers-Briggs Type Indicator. *Educational and Psychological Measurement*, *55*, 535-544.

Harvey, R. J., & Thomas, L. (1996). Using item response theory to score the Myers-Briggs Type Indicator: Rationale and research findings. *Journal of Psychological Type*, *37*, 16-60.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Johnson, D. A., & Saunders, D. R. (1990). Confirmatory factor analysis of the Myers-Briggs Type Indicator Expanded Analysis Report. *Educational and Psychological Measurement*, *50*, 561-571.

Laatsch, L., & Choca, J. (1994). Cluster-branching methodology for adaptive testing of the adaptive category test. *Psychological Assessment*, *6*, 1-7.

Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7).

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. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Maximum likelihood item analysis and test scoring-logistic models*. Chicago: Scientific Software International.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

Myers, I. B., & McCaulley, M. H. (1985). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Myers, I. B., McCaulley, M. H., Quenk, N., & Hammer, A. L. (1998). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator* (Rev. ed.). Palo Alto, CA: Consulting Psychologists Press.

Samejima, F. (1979). *A new family of models for the multiple-choice item* (Research Report 79-4 prepared under Office of Naval Research contract N00014-77-C-360, NR 150-402). Knoxville: Department of Psychology, University of Tennessee.

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computertized adaptive testing: From inquiry to operation*. Washington, DC: American Psychology Association.

Thissen, D., & Steinberg, L. (1985). A response model for multiple choice items. *Psychometrika*, *49*, 501-519.

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.
van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* Heidelberg, Germany: Springer-Verlag.
Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption Scale. *Journal of Personality and Social Psychology, 37*, 1051-1058.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.