Behaviormetrika
Vol.33, No.1, 2006, 75–102

ITEM RESPONSE THEORY: PAST PERFORMANCE, PRESENT DEVELOPMENTS, AND FUTURE EXPECTATIONS

Klaas Sijtsma* and Brian W. Junker**

We give a historical introduction to item response theory, which places the work of Thurstone, Lord, Guttman and Coombs in a present-day perspective. The general assumptions of modern item response theory, local independence and monotonicity of response functions, are discussed, followed by a general framework for estimating item response models. Six classes of well-known item response models and recent developments are discussed: (1) models for dichotomous item scores; (2) models for polytomous item scores; (3) nonparametric models; (4) unfolding models; (5) multidimensional models; and (6) models with restrictions on the parameters. Finally, it is noted that item response theory has evolved from unidimensional scaling of items and measurement of persons to data analysis tools for complicated research designs.

1. Historical context

The measurement of mental properties has been a long-lasting quest that originated in the 19th century and continues today. Significant sources to the development of the interest in measurement may be traced back to the 19th century in which French and German psychiatry emphasized mental illness and its influence on motor and sensory skills and cognitive behavior, German experimental psychology emphasized the standardization of the research in which the data are collected, and English genetics emphasized the importance of the measurement of individual differences using a well-defined methodology, expressing measurements as deviations from the group mean. In the early 20th century, Alfred Binet (Binet & Simon, 1905) was the first to actually develop and use what we would nowadays call a standardized intelligence test, and in the same era Charles Spearman (1904, 1910) developed the concepts and methodology of what would later be called classical test theory (CTT) and factor analysis.

1.1 Classical test theory

In psychometrics, CTT was the dominant statistical approach to testing data until Lord and Novick (1968) placed it in context with several other statistical theories of mental test scores, notably item response theory (IRT). To understand the underpinnings of

Key Words and Phrases: assumptions of IRT, cognitive diagnosis IRT models, historical perspective on IRT, item response theory, multidimensional IRT models, nonparametric IRT models, review of item response theory, unfolding IRT models

* Tilburg University, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, 5000 LE Tilburg, The Netherlands. E-mail: K.Sijtsma@uvt.nl

** Carnegie Mellon University, Pittsburgh PA 15213, USA. E-mail: brian@stat.cmu.edu

The authors wish to express their gratitude to Rhiannon L. Weaver and Elizabeth A. Ayers, both of the Department of Statistics at Carnegie Mellon University, whose comments on a previous draft of this manuscript were very helpful in clarifying the text.

CTT, note that measurements of mental properties such as test scores are the product of a complex interplay between the properties of the testee on the one hand, and the items administered to him/her and properties of the testing situation on the other hand. More precisely, the testee's cognitive processes are induced by the items (e.g., their difficulty level and the mental properties required to solve them), his/her own physical and mental shape (e.g., did the testee sleep well the night before he/she was tested? was his/her performance affected by an annoying cold?), and the physical conditions in which testing takes place (e.g., was the room well lighted? were other testees noisy? was the test instruction clear?). Fundamental to CTT is the idea that, if one were to repeat testing the same testee using the same test in the same testing situation, several of these factors (e.g., the testee's physical and mental well-being and the testing conditions in the room) are liable to exert an impact on the test performance and the resulting test score which may either increase or decrease the test score in an unpredictable way. Statistically, this means that a model describing test scores must contain a random error component; see Holland (1990) for other ways of accounting for the random error component in latent variable models.

Given this setup, CTT rests on the idea that, due to random error (denoted ε) an observable test score (denoted X_+) often is not the value representative of a testee's performance on the test (denoted T). Let an arbitrary testee be indexed v, then the CTT model is

$$X_{+v} = T_v + \varepsilon_v. \tag{1}$$

For a fixed testee, CTT assumes that the expected value of random error, ε_v, equals 0 across independent replications for the same examinee v; that is, $E(\varepsilon_v) = 0$. Then expectation across the testees in a population also is 0: $E_v[E(\varepsilon_v)] = 0$. In addition, CTT assumes that random error, ε, correlates 0 with any other variable, Y; that is, $\rho(\varepsilon, Y) = 0$. Finally, for a fixed testee, v, taking the expectation across replications of both sides of Equation 1 yields $T_v = E(X_{+v})$. This operational definition of the testee's representative test performance replaced the older platonic view of the true (hence, T_v) score as a stable person property to be revealed by an adequate measurement procedure.

Skipping many other important issues, we note that the main purpose of CTT is to determine the degree in which test scores are influenced by random error. This has lead to a multitude of methods for estimating the reliability of a test score, of which the lower bound called Cronbach's (1951) alpha is the most famous. In a particular population, a test has high reliability when random error, ε, has small variance relative to the variance of the true score, T. Cronbach, Gleser, Nanda, and Rajaratnam (1972) generalized CTT to allow one to decompose the variation of test scores into components attributable to various aspects of the response- and data-collection process, a form of analysis now known as generalizability theory. Examples of these aspects, or facets, of measurement include testees, items, testing occasions, clustering variables such as classrooms or schools, and of course random error. The resulting reliability definition then expresses the impact of random error relative to other sources of variation. CTT and its descendents continue to be a popular tool for researchers in many different fields, for constructing tests and

questionnaires.

Although CTT was the dominant statistical test model in the first half of the 20th century (e.g., see Guilford, 1936; and Gulliksen, 1950; for overviews), other developments were also taking place. In England, Walker (1931) set the stage for what later was to become known as Guttman (1944, 1950) scaling, by introducing the idea that if a testee can answer a harder question correctly then he or she should be able to answer easier questions on the same topic correctly as well. Walker also introduced the idea of an quantitative index measuring deviation from this deterministic model in real data (also, see Loevinger, 1948). This was a deterministic approach—without a random error component—to the analysis of data collected by a set of items that are assumed to measure one psychological property in common. A little earlier, in the 1920s, Thurstone (1927) developed his statistical measurement method of comparative judgment. Thurstone's work may be viewed as the most important probabilistic predecessor of item response theory (IRT).

1.2 Thurstone's model for comparative judgment

Like CTT, Thurstone's method used a random error for explaining test performance, and like IRT, response processes were defined as distributions of a continuous mental property. This continuous mental property can be imagined as a continuous dimension (say, a measurement rod), on which testees have measurement values indicating their relative level, and items are positioned as thresholds. Thurstone (1927; also see Michell, 1990; Torgerson, 1958) described the probability that stimulus j is preferred over stimulus k as a function of the dimension on which these two stimuli are compared and used the normal ogive to model behavioral variability. First, he hypothesized that the difference, t_{jk}, between stimuli j and k is governed by their mean difference, $\mu_j - \mu_k$, plus random error, ε_{jk}, so that $t_{jk} = \mu_j - \mu_k + \varepsilon_{jk}$. Then, he assumed that ε_{jk} follows a normal distribution, say, the standard normal. Thus, the probability that a respondent prefers j over k or, equivalently, that $t_{jk} > 0$, is

$$P(t_{jk} > 0) = P[\varepsilon_{jk} > -(\mu_j - \mu_k)] = \frac{1}{\sqrt{2\pi}} \int_{-(\mu_j - \mu_k)}^{\infty} \exp^{-t^2/2} dt. \tag{2}$$

Note that CTT did not model random error, ε, as a probabilistic function of the difference, $X_{+v} - T_v$ (see Equation 1), but instead chose to continue on the basis of assumptions about ε in a group (e.g. Lord and Novick, 1968, p.27) [i.e., $E(\varepsilon_v) = 0$ and $\rho(\varepsilon, Y) = 0$]. Thus, here lies an important difference with Thurstone's approach and, as we will see shortly, with modern IRT.

From our present day's perspective, the main contribution of Thurstone's model of comparative judgment, as Equation 2 is known, lies in the modeling of the random component in response behavior in such a way that estimation methods for the model parameters and methods for checking the goodness-of-fit of the model to the data could be developed. In contrast, older versions of CTT were tautologically defined: the existence of error, although plausible, was simply assumed, while error variance (and correlations of error with other variables) was not separately identifiable, in the statistical sense, in the model. As

a result, the assumptions could not be (dis-)confirmed with data unless different independent data sets from real replicated test administrations were available. We emphasize that later developments of generalizability theory and linear structural equations models offered many possibilities for estimating components of variance and other features of test scores as well as for goodness-of-fit testing; here we are merely pointing out historical differences.

1.3 Lord's normal ogive IRT model

IRT arose as an attempt to better formalize responses given by examinees to items from educational and psychological tests than had been possible thus far using CTT. Lord (1952) discussed the concept of the item characteristic curve or trace line (also, see Lazarsfeld, 1950; Tucker, 1946), now known as the item response function (IRF), to describe the relationship between the probability of a correct response to an item j and the latent variable, denoted θ, measured in common by a set of J dichotomously scores items. Let X_j be the response variable for item j ($j = 1, \ldots, J$), which is scored 1 if the answer was correct and 0 if the answer was incorrect. Then, $P_j(\theta) = P(X_j = 1|\theta)$ is the IRF. Lord (1952, p. 5) defined the IRF by means of the cumulative normal distribution function, such that, in our notation,

$$P_j(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_j} \exp^{-z^2/2} dz;\ z_j = a_j(\theta - b_j),\ a_j > 0. \tag{3}$$

Parameter b_j locates the IRF on the θ scale and is often interpreted as the difficulty parameter of the item. Parameter a_j determines the steepest positive slope of the normal ogive, which is located at $\theta = b_j$.

Equation 3 is essentially Thurstone's (1927) model of comparative judgment. This is seen most easily by redefining Lord's model as the comparison of a person v and an item j, the question being whether person v dominates item j ($\theta_v > b_j$), and assuming a standard normal error, ε_{vj}, to affect this comparison. That is, define $t_{vj} = a_j(\theta_v - b_j) + \varepsilon_{vj}$ and notice that, because of the symmetry of the normal distribution, integration from $-a_j(\theta_v - b_j)$ to ∞ yields the same result as from $-\infty$ to $a_j(\theta_v - b_j)$; then,

$$P_j(\theta) = P(t_{vj} > 0) = P[\varepsilon_{vj} > -a_j(\theta_v - b_j)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_j(\theta_v - b_j)} \exp^{-t^2/2} dt;\ a_j > 0. \tag{4}$$

Both Thurstone's and Lord's model relate probabilities of positive response to a difference in location parameters, $\mu_j - \mu_k$ in Equation 2 and $\theta_v - b_j$ in Equation 4. A difference between the models is that, unlike Equation 2, Equation 3 allows slopes of response functions to vary across the items, thus recognizing differences in the degree in which respondents with different θ values are differentially responsive to different items. However, this difference between the models seems to be technical more than basic.

An important difference may be the inclusion of a latent variable in Equation 3. As a result, Equation 3 compares a stimulus to a respondent and explicitly recognizes person variability on the dimension measured in common by the items. Here, individual differences in θ may be estimated from the model. Thurstone's model compares stimuli to one

another and does not include a latent variable. Thus, it may be considered a model for scaling stimuli.

However, Thurstone was also interested in measuring individual differences. Scaling and measuring were done as follows: In the first stage, respondents are asked to compare stimuli on a specific dimension; e.g., from each pair of politicians select that one that you think is the most persuasive. The ordering of the stimuli on the persuasiveness dimension is then estimated on the basis of Equation 2. In the second stage, treating the stimulus ordering as known, the ordering of the respondents on the persuasiveness scale is determined (Torgerson, 1958), for example by using the *Thurstone score*

$$T_v(w) = \frac{\sum_{j=1}^{J} w_j X_{vj}}{\sum_{j=1}^{J} X_{vj}} \tag{5}$$

where $w = (w_1, \ldots, w_J)$ is a set of weights reflecting the preference- or persuasiveness-order of the items, and X_{vj} is respondent v's response to item j. So, even though the latent variable is not part of the formal model in Equation 2, person ordering is one of the goals of the associated measurement procedure that is known as Thurstone scaling. Indeed, as we shall see below, modern developments in direct-response attitude and preference scaling often combine item- and person-scaling tasks, much as in IRT.

1.4 Deterministic models by Guttman and Coombs

Like Lord's model, Guttman's (1944, 1950) model compared item and person locations, but unlike Lord's model, it was deterministic in the sense that

$$\theta < b_j \Leftrightarrow P_j(\theta) = 0; \text{ and } \theta \geq b_j \Leftrightarrow P_j(\theta) = 1, \tag{6}$$

where b_j is a location parameter, analogous to b_j in Equation 4. Guttman's model predicts with complete certainty the item score as a function of the sign of $(\theta_v - b_j)$. Since real data are usually messier than the deterministic predictions of the this model, several coefficients were developed to express the distance of data from predictions based on Equation 6; a critical discussion is provided by Mokken (1971, chap. 2). The need to adapt Guttman's model to account for deviations from the perfect item ordering implied by Equation 6 was also at the basis of Mokken's (1971) approach to nonparametric, probabilistic IRT, to be discussed below in Section 3.3.

Another historical development was that of unfolding models for preference data (Coombs, 1964). Coombs' original deterministic model was similar to Guttman's, and may be stated as

$$P(X_j = 1|\theta) = \begin{cases} 1 & \text{if } |\theta - b_j| \leq d_j/2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where b_j is a location parameter and d_j is sometimes called the "latitude of acceptance". Coombs' model predicts with certainty that the respondent will endorse item j (say, a political statement or a brand of beer) if his/her θ (which may quantify political attitude, preference for bitterness or sweetness, etc.) is in an interval of length d_j centered at b_j,

and will not otherwise. The origin of the term "unfolding" (Coombs (1964) used it to describe a particular geometric metaphor for reconciling the conflicting preference orders given by different respondents for a set of stimuli) is hardly relevant anymore, and nowadays unfolding models, proximity models, ideal-point models, and models with unimodal IRFs, are all essentially the same thing, especially for dichotomous response data. The unfolding models that have an error component and thus are probabilistic are discussed later on. For the moment we concentrate on probabilistic models for dominance data, which are prevailing in modern IRT.

Thus far, this brief overview of main contributions to the development of mental measurement has emphasized the idea that a random measurement error is needed to describe the process of responding to an item with a reasonable degree of realism. Despite their lack of a mechanism for modeling the uncertainty that is typical of human response behavior, deterministic models such as those by Guttman and Coombs have been excellent vehicles for understanding the basic ideas of this response process. We will now continue outlining some of the key ideas of IRT.

2. Assumptions of IRT, and estimation

2.1 Assumptions of IRT and general model formulation.

Three key assumptions underlie most modern IRT modeling—and even IRT models that violate these assumptions do so in well-controlled ways. Letting x_j represent an arbitrary response to the j^{th} item (dichotomous or polytomous), we can write these cornerstones of the IRT approach as

- *Local independence (LI).* A d-dimensional vector of latent variables $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ exists, such that $P[X_1 = x_1, \ldots, X_J = x_j|\boldsymbol{\theta}] = \prod_{j=1}^{J} P[X_j = x_j|\boldsymbol{\theta}]$.
- *Monotonicity (M).* The functions $P[X_j = x_j|\boldsymbol{\theta}]$ satisfy the condition that for any ordered item score c, we have that $P[X_j > c|\boldsymbol{\theta}]$ is nondecreasing in each coordinate of $\boldsymbol{\theta}$, holding the other coordinates fixed.

When $d = 1$, we simply write θ instead of $\boldsymbol{\theta}$. This situation gets a special name,

- *Unidimensionality (U).* The dimension of $\boldsymbol{\theta}$ satisfying LI and M is $d = 1$.

and otherwise we call $\boldsymbol{\theta}$ *multidimensional*.

The properties M and U are already evident in Lord's Normal Ogive model, Equation 3: U holds by definition since θ there is unidimensional; and for dichotomous responses M boils down to the assumption that $P[X_j = 1|\theta]$ is nondecreasing in θ, which holds in Equation 4 because we assumed $a_j > 0$. These assumptions are intuitively appealing in educational testing, for example, where θ can be interpreted as quantifying some broad, test-relevant skill or set of skills, such as proficiency in school mathematics, and the test items are mathematics achievement items: the higher the proficiency θ, the more likely that an examinee should score well on each item. In addition, the assumptions M and U have proved useful in a wide variety of applications—involving hundreds of populations

and thousands of items—all across psychological and educational research, sociology, political science, medicine and marketing.

2.1.1 Local independence

The property LI is certainly computationally convenient, since it reduces the likelihood $P[X_1 = x_1, \ldots, X_J = x_j|\boldsymbol{\theta}]$ to a product of simpler terms that can be analyzed similarly to models for simple random samples in elementary statistics. However, it is also easily seen to be intuitively appealing. Indeed, if we let $\boldsymbol{x}_{(-j)}$ be the vector of $J-1$ item scores, omitting x_j, then LI implies

$$P[X_j = x_j \mid \boldsymbol{\theta}, \boldsymbol{X}_{(-j)} = \boldsymbol{x}_{(-j)}] = P[X_j = x_j \mid \boldsymbol{\theta}]. \tag{8}$$

That is, for the task of predicting the response on the j^{th} item, once we know $\boldsymbol{\theta}$, information from the other item responses is not helpful. In this sense $\boldsymbol{\theta}$ is a sufficient summary of the item responses.

Equation 8 also makes clear that LI is a strong condition that may not hold in all cases. For example, if the set of items is long and respondents learn while answering items, then Equation 8 is unlikely to hold. Also, if the respondent has special knowledge unmodeled by $\boldsymbol{\theta}$ regarding some items, or some items require special knowledge unmodeled by $\boldsymbol{\theta}$, then Equation 8 is again unlikely to hold. For this reason several alternatives to LI have been proposed. For example, Zhang and Stout (1999a) introduced the *weak LI (WLI)* condition,

$$Cov(X_j, X_k|\boldsymbol{\theta}) = 0, \tag{9}$$

which is implied by LI but, reversely, does not imply LI. Stout (1990) considered the even weaker *essential independence (EI)* condition, which can be written as

$$\lim_{J\to\infty} \binom{J}{2}^{-1} \sum_{1\le j<k\le J}\sum |Cov(X_j, X_k|\boldsymbol{\theta})| = 0. \tag{10}$$

There are at least two ways to look at LI. One is as a *desideratum* for measurement procedures. LI stipulates that the measurement procedure itself must not affect its outcome, such as would be caused by learning or other forms of development taking place while someone is being tested. Thus, given that LI is true, the items are modeled as stimuli that each function as a little experiment independent of the others; this is expressed by Equation 8. It also means that the statistical model (likelihood) for this measurement procedure reduces to a product of separate terms for each item, a form that is familiar and convenient for statistical computation. This is a strong assumption indeed, because human beings learn from experience and trying, say, 30 problems in an arithmetic test is likely to induce a learning process while doing this. Likewise, filling out a personality inventory is likely to induce the respondent to reflect upon himself in the process of rating the questions. Equation 8 may no longer hold under such circumstances.

Another way to look at LI is as a *criterion* for determining the dimensionality of the test data. Finding the dimensionality—the minimum number of latent traits necessary to explain the relationships between the items while possibly maintaining other restrictions

such as assumption M—is simple in principle: Add θ's until LI or its consequence, WLI (Equation 9), are satisfied as much as possible for the whole test. In practice, however, simply adding θ's is not a trivial thing to do, and may take different forms. For example, one approach selects items into clusters on the basis of the strength of their relationships with latent variables such that each cluster measures a different θ, while some items are possibly discarded altogether because they predominantly measure a unique θ (Molenaar & Sijtsma, 2000). Another approach may actually shift items around from one cluster to another until an estimate of the mean of conditional covariances as in WLI (Equation 9) is minimized for the particular data set (Stout et al., 1996; Zhang & Stout, 1999b). Parametric methods in particular may take the form of testing a null hypothesis that WLI holds for the whole test against the undirected alternative that WLI does not hold, and local tests can then be applied to find the item pairs responsible for misfit (e.g., Glas & Verhelst, 1995).

Of course, conditional independence (which is what we have called LI) is known in applied statistics, but these two approaches to LI are typical of IRT. LI as a measurement desideratum makes the test constructor aware of the importance of a controlled test procedure in which all unwanted influences should be eliminated beforehand or controlled afterwards through the use of auxiliary information collected on the respondents. LI as a criterion for dimensionality actually treats deviation from LI as a loss function, making the psychometrician aware of the inherent cognitive complexity of mental measurement; Stout's (1987, 1990) EI (Equation 10) assumption partially formalizes this idea by identifying the maximum deviation from LI that still allows estimation of a "dominant" unidimensional latent variable. Thus, unidimensional IRT models are little more than ideal data representations that may be fitted to data in a first attempt to learn about the composition of the data. Research into the dimensionality structure of the data may be an inevitable next step and multidimensional IRT models may be more important here than credited for thus far. No matter how one looks at LI, both visions stimulate the use and development of meaningful theories about the constructs to be measured in the process of test construction.

Local *dependence* may be inherent in a test procedure as in dynamic testing in which children are alternately trained and tested. The training involves abilities that do not become automatic, such as spatial learning ability, and the testing procedure monitors change in ability due to training. The development of the abilities causes individual differences in ability to become greater and may also cause the ability to become more complex by requiring more sub-abilities and skills to explain this variance. Embretson's (1991) multidimensional Rasch model for learning and change (MRMLC) formalizes these ideas. Jannarone's (1997) approach to local dependence more directly formalizes learning effects *during* testing, either due to training (e.g., by exposing correct answers to the previous items after the person has given his/her answer) or due to a natural process, as when dependence between items exists as a result of, for example, their reference to the same content domain as in verbal comprehension items that ask questions about the same short story; a general model of this type has been developed by Bradlow, Wainer and Wang (1999) for example.

2.1.2 Monotonicity

Similar to LI, one way to look at assumption M is as a *desideratum* needed to ascertain particular measurement properties for a test. Notice that although intuitively it is difficult to argue why an IRF would *not* be monotone —for example, why would the response probability of having an arithmetic item correct go down, even locally, with increasing ability?— logically there is no compelling reason why M should be true in real data. For example, although a regression curve fitted through the 0/1 item scores is very likely to have a positive slope, it is frequently found that the corresponding estimated IRFs significantly decrease at one or more intervals of θ. For particular abilities, an explanation may be that at consecutive θ intervals testees use different solution strategies that vary in degree of complexity and correctness (e.g., Bouwmeester, Sijtsma, & Vermunt, 2004). For example, for lower θ's the strategy may be simple and incorrect, for intermediate θ's it may be unnecessarily complex and partly correct, and for higher θ's the strategy may be simple and correct. Suppose that the items have multiple-choice format. It may be possible that, for some items but not all, the complex intermediate θ strategy leads testees more often astray to an incorrect answer than a simple, incorrect strategy that, by accident, produces several correct answers, just as flipping a coin would. The resulting IRF would show a shallow dip in the middle of the θ distribution.

In practice, for many items assumption M has been found to be reasonable in the regression sense just mentioned (curves have positive regression coefficients), but also many (small) deviations are found. Like assumption LI, assumption M is a strong assumption, in particular if it must hold for all J items in the test. Thus, for dichotomous items Stout (1987, 1990) proposed to replace M by *weak M* meaning that the test response function,

$$T(\boldsymbol{\theta}) = J^{-1} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}), \tag{11}$$

is increasing coordinate-wise in $\boldsymbol{\theta}$, for sufficiently large J. Weak M guarantees that there is enough information (in the sense of Equation 14 below) to estimate θ from the test scores. Stout (1990) showed that if weak M (Equation 11) and EI (Equation 10) together hold for a unidimensional θ then the total score $X_+ = \sum_{j=1}^{J} X_j$ consistently estimates (a transformation of) θ as $J \to \infty$, a result that was generalized to polytomous items by Junker (1991). In other words, also weaker forms of M, such as weak M, can be seen as *desiderata*, implying desirable measurement properties, in this case consistency.

However, if M is true for all J items, and LI and U hold, then the stochastic ordering of θ by means of total score X_+, in fact, by the unweighed sum score based on any subset of items from the set of J binary items, is true. That is, for any t and $x_{+v} < x_{+w}$, we have that

$$P(\theta > t | X_+ = x_{+v}) \leq P(\theta > t | X_+ = x_{+w}), \tag{12}$$

which implies that

$$E(\theta | X_+) \text{ monotone nondecreasing in } X_+.$$

This result includes special cases such as the Rasch (1960) model and the 3-parameter

logistic model (3PLM; e.g., Lord, 1980) but also Lord's normal ogive model (Equation 3). Hemker, Sijtsma, Molenaar, and Junker (1997) showed that Equation 12 also holds for the partial credit model (Masters, 1982) for ordered polytomous items, but not for other conventional polytomous IRT models; see Van der Ark (2005) for robustness results for many polytomous-item models.

In our experience, IRFs estimated from real data sets tend to be monotone, but local nonmonotonicities are not unusual. Few analyses proceed by assuming weak M only and dropping assumption M altogether, because in practice J is finite and often small, and in that case nonmonotone IRFs may lead to serious distortions in the ordering of persons by X_+ relative to θ. However, only allowing items that satisfy M in a test seems to be too strict, because nonmonotonicities are so common. In the subsection on nonparametric IRT models some methods for investigating assumption M are mentioned.

Another way to look at assumption M is as a *criterion* for the measurement quality of the items. It is common in IRT to express measurement quality by means of Fisher's information function. Let $P_j'(\theta)$ denote the first derivative of the IRF with respect to θ, then for dichotomous items Fisher's information function for item j is

$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{P_j(\theta)[1 - P_j(\theta)]}, \tag{13}$$

and, when LI holds, Fisher's information for the whole test is

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta). \tag{14}$$

Equation 13 gives the statistical information in X_j for every value of θ, and $I(\theta)^{-1/2}$ gives a lower bound on the standard error for estimating θ, which is achieved asymptotically for the maximum likelihood (ML) estimator and related methods, as $J \to \infty$. Clearly, the slopes of the IRFs at θ, $P_j'(\theta)$ (and, more specifically, the root-mean-square of all item slopes), determine measurement accuracy. Thus, although assumption M is important for interpreting IRT models, for measurement quality it matters more how steep the IRF is for values of θ where accurate measurement is important: near such θs, the IRF may be increasing or decreasing as long as it is steep, and the behavior of the IRF far from θ does not affect estimation accuracy near θ at all.

It follows from the discussion thus far, that for high measurement quality one needs to select items into a test that have high information values at the desired θ levels. Given that we would know how to first determine those levels and then how to find items that are accurate there, we are thus looking for steeply-sloped IRFs. Concepts like relative efficiency (Lord, 1980, chap. 6), item discrimination (e.g., see Equation 3), and item scalability (Mokken, 1997) help to find such items. See for example Van der Linden (2005) for a complete introduction to the test assembly problem.

2.2 Estimation of IRT models

Now we turn to general IRT modeling for the data matrix $\mathbf{X}_{N\times J}$ with entries x_{vj} that

we obtain when N subjects respond to J items, under the assumption that LI holds. Consider a testee sampled from a specific population by some process (simple random sampling, for example, or perhaps some more-complex process that is a combination of random sampling and administrative or social constraints); we will denote the distribution of $\boldsymbol{\theta}$ induced by the sampling process by $G(\boldsymbol{\theta})$. Given no other information than this, our best prediction—say, in the sense of minimizing squared-error loss—of any summary $f(\boldsymbol{\theta})$ of the sampled testee's $\boldsymbol{\theta}$ value is

$$E[f(\boldsymbol{\theta})] = \int_{\theta_1} \cdots \int_{\theta_d} f(\boldsymbol{\theta}) dG(\boldsymbol{\theta}).$$

Analogously, our best prediction of the respondent's probability of answering $\mathbf{x}_v = (x_{v1}, \ldots, x_{vJ})$ on the J items should be

$$P[X_{v1} = x_{v1}, \ldots, X_{vJ} = x_{vJ}] = \int_{\theta_1} \cdots \int_{\theta_d} P[X_{v1} = x_{v1}, \ldots, X_{vJ} = x_{vJ} | \boldsymbol{\theta}] dG(\boldsymbol{\theta}). \quad (15)$$

This is a basic building block for modeling IRT data. If we consider the data matrix $\mathbf{X}_{N\times J}$, and we assume that respondents are sampled i.i.d. (independently and identically distributed) from $G(\boldsymbol{\theta})$, we obtain the model

$$P[\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}] = \prod_{v=1}^{N} P[X_{v1} = x_{v1}, \ldots, X_{vJ} = x_{vJ}] \quad (16)$$

for $\mathbf{X}_{N\times J}$. This i.i.d. assumption is sometimes called *experimental independence* in IRT work, and might hold for example if the respondents have a common background relevant to the items (so that there are no unmodeled variance components among them) and did not collaborate in producing item responses.

The model for $\mathbf{X}_{N\times J}$ in Equation 16 can be elaborated, using LI and the integral representation in Equation 15, to read

$$P[\mathbf{X}_{N\times J} = \mathbf{x}_{N\times J}] = \prod_{v=1}^{N} \int_{\theta_1} \cdots \int_{\theta_d} \prod_{j=1}^{J} P[X_{vj} = x_{vj} | \boldsymbol{\theta}] dG(\boldsymbol{\theta}). \quad (17)$$

If the model for $P[X_{vj} = x_{vj}|\boldsymbol{\theta}]$ is the normal ogive model (Equation 3) for example, then the probability on the left in Equation 17 is a function of $2J$ parameters ($a_1, \ldots, a_J, b_1, \ldots, b_J$) and these—as well as a fixed number of parameters of the distribution $G(\theta)$—can be estimated by ML. The approach to estimation and inference based on Equation 17 is called the *marginal maximum likelihood (MML)* approach, and is widely favored because it generally gives consistent estimates for the item parameters as the number N of respondents grows. Such an approach might be followed by empirical Bayes inferences about individual examinees' θs (see e.g. Bock & Mislevy, 1982).

It should be noted that there are other ways of arriving at a basis for inference similar to Equation 17. For example, we may wish to jointly estimate the item parameters ($a_1, \ldots, a_J, b_1, \ldots, b_J$) and each respondent's latent variables $\boldsymbol{\theta}_v$, $v = 1, \ldots, N$. This leads to a different likelihood,

$$P[\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N] = \prod_{v=1}^{N} \prod_{j=1}^{J} P[X_{vj} = x_{vj}|\boldsymbol{\theta}], \tag{18}$$

and an estimation method called *joint maximum likelihood (JML)*. If the model of Equation 3 is used in Equation 18, we would be estimating $2J + N$ parameters. JML is viewed as generally less attractive than MML, since it can be shown (Neyman & Scott, 1948; Andersen, 1980) that estimates of item parameters, for example, can be inconsistent with these parameters' true values as we obtain more and more information—i.e., as N and $J \rightarrow \infty$—unless N and J are carefully controlled (Haberman, 1977; Douglas, 1997).

The integration in Equation 17 obviates this inconsistency problem, since it effectively eliminates $\boldsymbol{\theta}$ from the model for estimating item parameters to the observable data $\mathbf{X}_{N\times J}$. In some models in which $P[X_{vj} = x_{vj}|\boldsymbol{\theta}]$ is a member of the exponential family of distributions, $\boldsymbol{\theta}$ can be eliminated by conditioning on sufficient statistics S_v for each $\boldsymbol{\theta}_v$. Thus the JML likelihood is transformed into

$$P[\boldsymbol{X} = \boldsymbol{x}|S_1, \ldots, S_N],$$

which is only a function of the item parameters. The method of estimation based on this conditional likelihood is called *conditional maximum likelihood (CML)* and is well known for estimating parameters in the Rasch (1960) model, where the sufficient statistic for θ_v is respondent v's total score, $S_v = X_{+v} = \sum_{j=1}^{J} x_{vj}$. Andersen (1980) and others [e.g., see Holland (1990), for a useful review] showed that CML estimates of item parameters are consistent with the true values, as J grows, just as MML estimates are.

Finally, a Bayesian model-building approach also leads to a form similar to Equation 17. In this approach, items are viewed as exchangeable with each other, leading to LI conditional on θ, and to a pre-posterior model formally equivalent to Equation 15, conditional on the item parameters. Testees are then also viewed as exchangeable with one another, leading to a model of the form Equation 16. Finally, if $G(\boldsymbol{\theta})$ is an informative prior distribution for $\boldsymbol{\theta}$ (perhaps based on knowledge of the sampling process producing respondents) and we place flat noninformative priors on the item parameters, then posterior mode ("maximum a-posteriori", or *MAP*, in much IRT literature) estimates of item parameters are identical to those obtained by maximizing Equation 17. Although in the present context it seems like the Bayesian approach is nothing more than a trivial restatement of the assumptions leading to Equation 17 we will see below that the Bayesian perspective is a powerful one that has driven much of the modern expansion of IRT into a broad toolbox for item response modeling in many behavioral and social science settings.

Some restrictions along the lines of LI, M and U are needed to give the model in Equation 17 "bite" with data, and therefore strength as an explanatory or predictive tool. Although there is much latitude to weaken these assumptions in various ways, no one of them can be completely dropped, or the model will simply re-express the observed distribution of the data—maximizing capitalization on chance and minimizing explanatory or predictive power. For example, Suppes and Zanotti (1981; also Holland & Rosenbaum, 1986; Junker, 1993) have shown that the structure in Equation 17 does not restrict the distribution of the data, unless the response functions and/or the distribution of $\boldsymbol{\theta}$ are

restricted. This is what IRT has done: Assuming LI, IRT concentrates mostly on the response functions and finds appropriate definitions for them in an attempt to explain the simultaneous relationships between the J items through Equation 17. The distribution of $G(\boldsymbol{\theta})$ may be restricted primarily to facilitate the estimation of IRT model parameters, a common practice in Bayesian and MML approaches to estimation.

3. Some well-known classes of IRT models

3.1 IRT models for dichotomous items

For dichotomous item scores for correct/incorrect or agree/disagree responses, many IRT models have been defined based on assumptions LI and U, to which a parametric version of assumption M is added. Due to their computational merits, logistic models have gained great popularity. An example is the three-parameter logistic model (3PLM; Birnbaum, 1968), defined as

$$P_j(\theta) = \gamma_j + (1 - \gamma_j)\frac{\exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]},\ \alpha_j > 0. \tag{19}$$

In Equation 19, parameter γ_j is the lower asymptote of the IRF, that gives the probability of, for example, a correct answer for low θ's. This parameter is sometimes interpreted as a guessing parameter. Parameter δ_j is the location or difficulty, comparable to b_j in Equation 3. Parameter α_j determines the steepest slope of the IRF, comparable to a_j in Equation 3. This occurs when $\theta = \delta_j$; then $P_j(\theta) = \frac{1+\gamma_j}{2}$ and $P_j'(\theta) = \frac{\alpha_j(1-\gamma_j)}{4}$.

The 3PLM is suited in particular for fitting data from multiple-choice items that vary in difficulty and discrimination and are likely to induce guessing in low-ability examinees. Its parameters can be estimated using ML methods. If in Equation 19 $\gamma_j = 0$, the 3PLM reduces to the 2-parameter logistic model (2PLM; Birnbaum, 1968), which is almost identical to the 2-parameter normal-ogive in Equation 3. The 1-parameter logistic model (1PLM) or Rasch model (Fischer & Molenaar, 1995; Rasch, 1960) sets $\gamma_j = 0$ and $\alpha_j = 1$ in Equation 19. Notice that such models may be interpreted as probabilistic versions of the deterministic Guttman model (Equation 6).

3.2 IRT models for polytomous items

IRT models have been defined for items with more than two possible but *nominal* scores, such as in Bock's (1972) nominal response model and Thissen and Steinberg's (1984) response model for multiple-choice items. Such models are convenient when different answer categories have to be distinguished that do not have an a priori order. Other IRT models have been defined for *ordered* polytomous item scores; that is, items for which $X_j = 0, \ldots, m$, typical of rating scales in personality inventories and attitude questionnaires. Given assumptions LI and U, three classes of polytomous IRT models have been proposed (Mellenbergh, 1995). Hemker, Van der Ark, and Sijtsma (2001) provide a Venn diagram that shows how these three classes and their member models are hierarchically related.

One such general class consists of the cumulative probability models. Such models are typically based on the monotonicity of conditional response probability $G_{jx}(\theta) = P(X_j \geq x|\theta)$. This probability is the item step response function (ISRF). A well known representative of this class is the homogeneous case of the graded response model (Samejima, 1969, 1997), that has a constant slope parameter for each ISRF from the same item, and a location parameter that varies freely, so that

$$G_{jx}(\theta) = \frac{\exp[\alpha_j(\theta - \delta_{jx})]}{1 + \exp[\alpha_j(\theta - \delta_{jx})]}, x > 0, \alpha_j > 0.$$

Notice that this response function is equivalent to that of the 2PLM, but that the difference lies in the item score that is modeled: polytomous $X_j \geq x$ in the graded response model and binary $X_j = 1$ in the 2PLM, and that they coincide when $m = 1$. Cumulative probability models are sometimes associated with data stemming from a respondent's global assessment of the rating scale and the consecutive choice of a response option from all available options (Van Engelenburg, 1997).

The other classes are the continuation ratio models (see, e.g., Hemker, et al., 2001) and the adjacent category models (Hemker, et al., 1997; Thissen & Steinberg, 1986). Continuation ratio models define response probabilities $M_{jx}(\theta) = P(X_j \geq x|X_j \geq x - 1; \theta)$. Hemker et al. (2001) argue that such models formalize the performance on a sequence of subtasks of which the first x were mastered and as of $x + 1$ the others were failed; also see Tutz (1990) and Samejima (1972, chap. 4). Adjacent category models define response functions $F_{jx}(\theta) = P(X_j = x|X_j \in \{x - 1, x\}; \theta)$. Like the continuation ratio models, adjacent category models "look at" the response process as a sequence of subtasks but unlike these models define a subtask in isolation of the others. This is formalized by the conditioning on scores $x - 1$ and x only, and it means that the subtasks do not have a fixed order but that each can be solved or mastered independent of the others. The partial credit model (Masters, 1982), which defines $F_{jx}(\theta)$ as a 1PLM, is perhaps the best known model from this class.

Many of the models discussed above are reviewed and extended, often by the models' originators, in various chapters of the monograph by Van der Linden and Hambleton (1997).

One area in which polytomous items arise naturally is in the rating of extended responses by trained raters or judges. When extended response items are scored by more than one rater, the repeated ratings allow for the consideration of individual rater bias and variability in estimating student proficiency. Several hierarchical models based on IRT have been recently introduced to model such effects. Patz, Junker, Johnson and Mariano (2002) developed a *hierarchical rater model* to accommodate additional dependence between ratings of different examinees by the same rater. The hierarchical rater model assumes that rating X_{ijr} given to examinee i on item j by rater r is a noisy, and perhaps biased, version of the "ideal rating" ξ_{ij} that the examinee would get from a "perfect rater". The ideal ratings ξ_{ij} are modeled in the hierarchical rater model using a polytomous IRT model such as the partial credit model. The observed ratings X_{ijr} are then modeled conditional on the ideal ratings ξ_{ij}; for example, Patz et al. (2002) specify a

unimodal response function for raters' ratings, given the "ideal rating", of the form

$$P[X_{ijr} = x|\xi_{ij} = \xi] \propto \exp\left\{-\frac{1}{2\psi_r^2}[x - (\xi + \phi_r)]^2\right\},$$

where ϕ_r is the bias of rater r across all examinees and rated items, and ψ_r is a measure of the rater's uncertainty or unreliability in rating. This specification of the hierarchical rater model essentially identifies it as a member of the class of multilevel models (see e.g. Gelman et al., 2004), which includes variance components models as well as the hierarchical linear model. The connection with variance components models also allows us to see deep connections between the hierarchical rater model and generalizability theory models; see Patz et al. (2002) for details.

The hierarchical rater model has been successfully applied to paper-and-pencil rating of items on one large statewide assessment in the USA, and to a comparison of "modes of rating" (computer image-based vs. paper-and-pencil) in another statewide exam that included both rated extended-response items. A partial review of some other approaches to modeling multiply-rated test items, as well as an extension of the basic hierarchical rater model of Patz et al. (2002) to accommodate covariates to help explain heterogeneous rating behaviors, may be found in Mariano and Junker (2005).

3.3 Nonparametric IRT models

Nonparametric IRT seeks to relax the assumptions of LI, M, and U, while maintaining important measurement properties such as the ordinal scale for persons (Junker, 2001; Stout, 1990, 2002). One such relaxation is that nonparametric IRT models refrain from a parametric definition of the response function, as is typical of IRT models discussed thus far. For example, for dichotomous items the monotone homogeneity model (MHM; Mokken, 1971; also see Meredith, 1965) assumes LI, M, and U as in Section 2, but no parametric form for the IRFs.

Notice that, for example, the 3PLM is a special case of the MHM because it restricts assumption M by means of the logistic IRF in Equation 19. Reasons why the 3PLM may not fit data are that high-ability examinees have response probabilities that are smaller than 1 or the relationship between the item score and the latent variable does not have a smooth logistic appearance. Thus, in such cases one needs a more flexible model that allows for the possibility that the upper asymptote of the response functions is smaller than 1 and the curve can take any form as long as it is monotone.

Assumption M has been investigated in real data by means of its observable consequence, manifest monotonicity (e.g., Junker & Sijtsma, 2000): Let $R_{(-j)} = \sum_{k\neq j} X_k$ be the total score on the $J-1$ dichotomous items excepting item j; this is called the *rest-score* for item j. Then monotonicity of $P_j(\theta)$, together with LI and U, implies

$$P[X_j = 1|R_{(-j)} = r] \text{ non-decreasing in } r, r = 0, \ldots, J-1.$$

Manifest monotonicity does not hold when conditioning is on X_+ including X_j, nor does it hold for polytomous items; Junker (1996) has suggested another methodology for the

latter case.

For dichotomous items, $P[X_j = 1|R_{(-j)} = r]$ is the basis for estimating the item's IRF by estimating response probabilities for each discrete value of r using what is known in nonparametric regression as binning (Junker & Sijtsma, 2000). Karabatsos and Sheu (2004) proposed a Bayesian approach using Markov Chain Monte Carlo simulation to evaluating assumption M for J items simultaneously. This procedure also gives information about item fit. Alternatively, kernel smoothing methods may be used to obtain a continuous estimate of the IRF, the ISRFs (for polytomous items), and the so-called option response curves, $P(X_j = x_j|\theta)$, with X_j nominal, representing the options of a multiple-choice item (Ramsay, 1991). Jack-knife procedures may be used to estimate confidence envelopes (Emons, Sijtsma, & Meijer, 2004). Rossi, Wang, and Ramsay (2002) proposed a methodology that uses EM likelihood estimation to obtain the logit transformation of the IRF, denoted $\lambda_j(\theta)$, by means of a linear combination of polynomials, chosen by the researcher and used to approximate adjacent segments of $\lambda_j(\theta)$. Each polynomial has a weight which is estimated from the data, and which controls the smoothness of $\hat{\lambda}_j(\theta)$. As with kernel smoothing, a very irregular curve may actually show much sampling error and a very smooth curve may mask systematic and interesting phenomena that are useful to diagnose the item.

Although they are based on weaker assumptions than parametric models, nonparametric IRT models have desirable measurement properties, such as $P(\theta > t|X_+ = x_+)$ is nondecreasing in x_+ (Equation 12; for dichotomous items only) and X_+ is a consistent ordinal estimator of θ under relaxed versions of LI, M and U (both for dichotomous and polytomous items). The fit of nonparametric models is often evaluated by finding the dimensionality of the data such that weak LI is satisfied (e.g., Stout et al, 1996), and by estimating, for example, the IRFs and check whether assumption M holds (e.g., Ramsay, 1991; Junker & Sijtsma, 2000).

Powerful tools for this fit investigation are based on *conditional association* (Holland & Rosenbaum, 1986), which says that, for any partition $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$; any nondecreasing functions $n_1()$ and $n_2()$, and any arbitrary function $m()$, LI, M, and U imply

$$Cov[n_1(\mathbf{Y}), n_2(\mathbf{Y})|m(\mathbf{Z}) = z] \geq 0, \text{ for all } z. \tag{20}$$

Judicious choice of the functions n_1, n_2 and m will readily suggest many meaningful ways of checking for the general class of models based on LI, M, and U. We mention two important ones that form the basis of methods for dimensionality investigation.

First, letting $m(\boldsymbol{Z})$ be the function identically equal to zero, Equation 20 implies that

$$Cov(X_j, X_k) \geq 0, \text{ all pairs } j, k; j \neq k. \tag{21}$$

Mokken (1971; Sijtsma & Molenaar, 2002) proposed a procedure that uses a scalability coefficient that incorporates Equation 21 for all item pairs as a loss function for item selection, and constructs item clusters that tend to be U while requiring items to have relatively steep (positive) slopes to be admitted. Minimum admissible steepness is controlled by the researcher. As is typical of trade-offs, higher demands with respect to slopes will

result is shorter unidimensional scales.

Second, we may condition on a kind of "rest score" for two items j and k, $R_{(-j,-k)} = \sum_{h \neq j,k} X_h$. In this case, Equation 20 implies that

$$Cov(X_j, X_k | R_{(-j,-k)} = r) \geq 0, \text{ all } j, k; j \neq k; \text{ all } r = 0, 1, \ldots, J-2. \tag{22}$$

Thus, in the subgroup of respondents that have the same rest score r, the covariance between items j and k must be nonnegative if they trigger responses that are governed by the same θ. The interesting part is that Zhang and Stout (1999a, b) have shown how the sign behavior of the conditional covariance in Equation 22 is related to the dimensionality of an item set. This sign behavior is the basis of a genetic algorithm that selects items into clusters within which WLI (Equation 9) holds as well as possible for the given data.

Similar work has been done for polytomous response scores. For example, Van der Ark, Hemker, and Sijtsma (2002) defined nonparametric versions of each of the three classes of polytomous IRT models, showed that each was the most general representative of its class and also proved that they were hierarchically ordered. In particular, the nonparametric version of the graded response model is the most general model for polytomous item scores among the existing models based on LI, U, and M for the ISRFs, and all other polytomous models, nonparametric and parametric, are special cases of it.

Not only does knowledge like this bring structure among the plethora of IRT models but it also suggests an order in which models can be fitted to data, beginning with either the most general and when it fits, continuing with fitting more specific models until misfit is obtained. Another methodology could start at the other end, using a restrictive model and when it does not fit, use a slightly less restrictive model, and so on, until a fitting model is found. Also, considerations about the response process could play a role in the selection of models. For example, if an item can be solved by solving a number of part problems in an arbitrary order an adjacent category model may be selected for data analysis.

3.4 Unfolding IRT models

Probabilistic versions of the Coombs unfolding model introduced in Equation 7 for binary preference scores (e.g. testees' preferences for brands of beer, people from other countries, or politicians, on the basis of dimensions such as bitterness, trustworthiness, and conservatism) have also been developed. IRT models for such direct-response attitude or preference data generally employ unimodal, rather than monotone, IRFs. We will call such models *unfolding IRT models*, though as mentioned in Section 1.4, the connection with Coombs' original unfolding idea is now rather tenuous.

Although unfolding IRT models have been around for years (e.g., Davison, 1977), it is only relatively recently that a close connection between these unimodal models (also known as proximity models) and conventional monotone IRT models (also known as dominance models) has been made, through a missing data process (Andrich & Luo, 1993; Verhelst & Verstralen, 1993). For example, the hyperbolic cosine model for dichotomous attitude responses ($X_j = 1$ for "agree"; $X_j = 0$ for "disagree"),

$$P_j(\theta) \;=\; P[X_j = 1 \mid \theta] = \frac{e^{\gamma_j}}{e^{\gamma_j} + \cosh(\theta - \beta_j)} = \frac{\exp\{\theta - \beta_j + \gamma_j\}}{1 + \exp\{\theta - \beta_j + \gamma_j\} + \exp\{2(\theta - \beta_j)\}} \;, \tag{23}$$

in which β_j is the location on the preference or attitude scale of persons most likely to endorse item j, and γ_j is maximum log-odds of endorsement of the item, can be viewed as the observed-data model corresponding to a complete-data model based on a trichotomous item response model (i.e., the partial credit model of Masters, 1982),

$$\begin{aligned}
R_{j0}(\theta) \;&=\; P[\xi_j = 0 \mid \theta] = \frac{1}{1 + \exp\{\theta - \beta_j + \gamma_j\} + \exp\{2(\theta - \beta_j)\}} \\
R_{j1}(\theta) \;&=\; P[\xi_j = 1 \mid \theta] = \frac{\exp\{\theta - \beta_j + \gamma_j\}}{1 + \exp\{\theta - \beta_j + \gamma_j\} + \exp\{2(\theta - \beta_j)\}} \\
R_{j2}(\theta) \;&=\; P[\xi_j = 2 \mid \theta] = \frac{\exp\{2(\theta - \beta_j)\}}{1 + \exp\{\theta - \beta_j + \gamma_j\} + \exp\{2(\theta - \beta_j)\}} \;,
\end{aligned}$$

where the complete data (or equivalently the "latent response", as in Maris, 1995) ξ_j is coded as

$$\begin{aligned}
&\xi_j = 0 \quad \text{if } \theta - \beta_j \;\ll\; 0 \quad \text{(i.e., "disagree from below")} \\
&\xi_j = 1 \quad \text{if } \theta - \beta_j \;\approx\; 0 \quad \text{(i.e., "agree")} \\
&\xi_j = 2 \quad \text{if } \theta - \beta_j \;\gg\; 0 \quad \text{(i.e., "disagree from above")} \;,
\end{aligned}$$

in which the distinction between "disagree from above" ($\xi_j = 2$) and "disagree from below" ($\xi_j = 0$) has been lost. Another such parametric family has been developed by Roberts, Donoghue and Laughlin (2000). Recently Johnson and Junker (2003) generalized this missing data idea by connecting it with importance sampling, and used it to develop computational Bayesian estimation methods for a large class of parametric unfolding IRT models.

Post (1992) developed a nonparametric approach to scaling with unfolding IRT models based on probability inequalities, stochastic ordering, and related ideas. A key idea is that $P[X_i = 1|X_j = 1]$ should increase as the modes of the IRFs for items i and j get closer together; this is a kind of "manifest unimodality" property. Post shows that this manifest unimodality property follows from suitable stochastic ordering properties on the IRFs themselves, which are satisfied, for example, by the model in Equation 23 when the γ_j's are all equal. Johnson (2005) re-examined Post's (1992) ground-breaking approach and connected it to a nonparametric estimation theory for unfolding models based on the work of Stout (1990), Ramsay (1991) and Hemker et al. (1997). For example, Johnson establishes monotonicity and consistency properties of the Thurstone score in Equation 5 under a set of assumptions similar to Hemker et al.'s (1997) and Stout's (1990), and explores estimation of IRFs via nonparametric regression of item responses onto the Thurstone score, similar to Ramsay's (1991) approach.

3.5 Multidimensional IRT models

Data obtained from items that require the same ability can be displayed in a two-

dimensional space, with the latent variable on the abscissa and the response probability on the ordinate. For such data, an IRT model assuming U (unidimensional θ) is a likely candidate to fit. If different items require different or partly different abilities, a higher-dimensional representation of the data is needed and models assuming U will probably fail to fit in a satisfactory way. Then, an IRT model that postulates d latent variables may be used to analyze the data.

Research in multidimensional IRT models has concentrated on additive and conjunctive combinations of multiple traits to produce probabilities of response. Additive models, known as *compensatory* models in much of the literature, replace the unidimensional latent trait θ in a parametric model such as the 2PL model, with an item-specific, known (e.g., Adams, Wilson, & Wang, 1997; Embretson, 1991; Kelderman & Rijkes, 1994; and Stegelmann, 1983) or unknown (e.g., Fraser & MacDonald, 1988; Muraki & Carlson, 1995; Reckase, 1985; and Wilson, Wood, & Gibbons, 1983) linear combination of components $\alpha_{j1}\theta_1 + \cdots + \alpha_{jd}\theta_d$ of a d-dimensional latent variable vector. For example, Reckase's (1997) linear logistic multidimensional model incorporates these parameters and is defined as

$$P_j(\boldsymbol{\theta}) = \gamma_j + (1-\gamma_j)\frac{\exp(\boldsymbol{\alpha}_j'\boldsymbol{\theta} - \delta_j)}{1 + \exp(\boldsymbol{\alpha_j}'\boldsymbol{\theta} - \delta_j)},$$

where the location parameter δ_j is related (but not identical) to the distance of the origin of the space to the point of steepest slope in the direction from the origin, and the γ_j parameter represents the probability of a correct answer when the θ's are very low. Note that the discrimination vector $\boldsymbol{\alpha}$ controls the slope (and hence the information for estimation) of the item's IRF in each coordinate direction: For example, for $\boldsymbol{\theta} = (\theta_1, \theta_2)$, if responses to item j are driven more by θ_2 than by θ_1, the slope (α_{j2}) of the manifold is steeper in the θ_2 direction than that (α_{j1}) in the θ_1 direction. Béguin and Glas (2001) survey the area well and give an MCMC algorithm for estimating these models; Gibbons and Hedeker (1997) pursue related developments in biostatistical and psychiatric applications. De Boeck and Wilson (2004) have organized methodology for exploring these and other IRT models within the SAS statistical package.

Conjunctive models are often referred to as *noncompensatory* or *componential* models in the literature. These models combine unidimensional models for components of response multiplicatively, so that $P(X_{vj} = 1|\theta_{v1}, \ldots, \theta_{vd}) = \prod_{\ell=1}^{d} P_{j\ell}(\theta_{v\ell})$ where $P_{j\ell}(\theta_{v\ell})$ are parametric unidimensional response functions for binary scores. Usually the $P_{j\ell}(\theta_{v\ell})$s represent skills or subtasks all of which must be performed correctly in order to generate a correct response to the item itself. Janssen and De Boeck (1997) give a typical application.

Compensatory structures are attractive because of their conceptual similarity to factor analysis models. They have been very successful in aiding the understanding of how student responses can be sensitive to major content and skill components of items, and in aiding parallel test construction when the underlying response behavior is multidimensional (e.g., Ackerman, 1994). Noncompensatory models are largely motivated from a desire to model cognitive aspects of item response; see for example Junker and Sijtsma (2001). Embretson (1997) reviewed blends of these two approaches (her general component latent trait models; GLTM).

3.6 IRT models with restrictions on the item parameters

An important early model of the cognitive processes that lead to an item score is the linear logistic test model (LLTM; Fischer, 1973; Scheiblechner, 1972). The LLTM assumes that the difficulty parameter, δ_j, of the Rasch model is a linear combination of K basic parameters, η_k, with weights Q_{jk}, for the difficulty of a task characteristic or a subtask in a solution strategy: $\delta_j = \sum_{k=1}^{K} Q_{jk}\eta_k + c$, where c is a normalization constant for the item parameters. The choice of the number of basic parameters and the item difficulty structure expressed by the weights and collected in a weight matrix $\mathbf{Q}_{J\times K}$, together constitute a hypothesis that is tested by fitting the LLTM to the 1/0 scores for correct/incorrect answers to the J test items.

Other models have been proposed that, for example, posit multiple latent variables (Kelderman & Rijkes, 1994), strategy shift from one item to the next (Rijkes, 1996; Verhelst & Mislevy, 1990), and a multiplicative structure on the response probability, $P_j(\theta)$ (Embretson, 1997; Maris, 1995).

Models for cognitive diagnosis often combine features of multidimensional IRT models with features of IRT models with restrictions on the item parameters. Suppose a domain requires in total K different skills, and we code for each respondent $\theta_{vk} = 1$ if respondent v has skill k and $\theta_{vk} = 0$ otherwise. As in the LLTM, we define $Q_{jk} = 1$ if item j requires skill k and $Q_{jk} = 0$ otherwise. Two simple conjunctive models for cognitive diagnosis were considered by Junker and Sijtsma (2001), the *NIDA* (noisy inputs, deterministic "and" gate) model,

$$P_j(\theta_v) = P(X_j = 1|\theta_v) = \prod_{k=1}^{K} \left[(1 - s_k)^{\theta_{vk}} g_k^{1-\theta_{vk}}\right]^{Q_{jk}} ,$$

where $s_k = P[(\text{slipping when applying skill } k) \mid \theta_{vk} = 1]$ and $g_k = P[(\text{succeeding where skill } k \text{ is needed}) \mid \theta_{vk} = 0]$, and the *DINA* (deterministic inputs, noisy "and" gate) model

$$P_j(\theta_v) = P(X_j = 1|\theta_v) = (1 - s_j)^{\prod_k \theta_{vk}^{Q_{jk}}} g_j^{1-\prod_k \theta_{vk}^{Q_{jk}}}$$

where now s_j and g_j play a similar role for the entire item rather than for each skill individually.

More elaborate versions of these conjunctive discrete-skills models have been developed by others (e.g., DiBello, Stout, & Roussos, 1995; Haertel, 1989; Maris, 1999; and Tatsuoka, 1995); and the NIDA and DINA models themselves have been extended to accommodate common variation among the skills being acquired (De la Torre & Douglas, 2004). A compensatory discrete-skills model was considered by Weaver and Junker (2003). Focusing on $\boldsymbol{\theta}_v = (\theta_{v1}, \ldots, \theta_{vK})$ these models have the form of multidimensional IRT models, with dimension $d = K$. Focusing on the restrictions on item response probability imposed by the Q_{jk}, they have the form of IRT (or latent class) models with restrictions on the item parameters. Junker (1999) provides an extended comparison of these and other models for cognitively-relevant assessment.

4. Discussion

In this paper we have sketched the historical antecedents of IRT, the models that have formed the core of IRT for the past 30 years or so, and some extensions that have occupied the interests of IRT researchers in recent years.

One of the earliest antecedents of IRT, classical test theory, is primarily a conceptual model that provides a simple decomposition of test scores into a reliable or "true score" component and an unreliable random error component; in this sense, CTT is a kind of variance components model. In the simplest form of CTT the true score and random error components are not identifiable. However, a variety of heuristic methods have been developed to estimate or bound the true score and random error components of the model, and so CTT is still widely used today as a handy and simple guide to exploring trait-relevant (true-score) variation vs. trait irrelevant (random error) variation in the total score of a test: A test is considered reliable and generalizable if little of the variation in test scores can be attributed to random error.

In one direction, CTT was refined and generalized to incorporate a variety of covariate information, treating some covariates as additional sources of stochastic variation and others as sources of fixed differences between scores. The resulting family of models, which goes under the name generalizability theory, allows for quite general mixed-effects linear modeling of test scores, partitioning variation of the test scores into components attributable to various aspects of the response- and data-collection process, as well as multidimensional response data. Although CTT and generalizability theory can be used to assess the quality of data collection, for example, expressed as a fraction of the total variation of scores due to noise or nuisance facets, they do not by themselves provide efficient model-based tools for estimating latent variables.

In another direction, CTT was generalized, first, to accommodate discrete-response data, as in Thurstone's model of comparative judgment, and later, as in Lord's Normal Ogive model to explicitly incorporate a common latent variable across all responses to items on the same test or questionnaire. At the same time other authors were developing related parametric probabilistic models for responses to test items (Rasch, 1960), as well as deterministic models for dominance (Guttman, 1944, 1950) and proximity/unfolding items (Coombs, 1964). These threads were drawn together into a coherent family of probabilistic models for measurement, called Item Response Theory, in the early 1960s.

IRT was certainly a conceptually successful model, because it provided parameters to estimate "major features" of test questions as well as examinee proficiencies. It was also a fantastically successful model on practical grounds, since, with the inexorable advance of cheap computing power in the 40 years from 1960 to 2000, IRT models could be applied to the vast quantities of primarily educational testing data being generated by companies like ETS in the United States and CITO in the Netherlands. Although the psychological model underlying IRT was not deep, it was adequate for many purposes of large-scale testing, including

- *Scaling:* Pre-testing new items to make sure that they cohere with existing test items

in the sense that LI, M, $d = 1$ still hold;

- *Scoring:* IRT-model-based scores, computed using ML or a similar method, offer more efficient, finer-grained scores of examinees than simple number-correct scores;
- *Equating:* IRT modeling, which separates person-effects from item-effects provides a formal methodology for attempting to adjust scores on different tests (of the same construct) for item-effects (e.g., item difficulty), so that they are comparable. The same methodology also allows one to pre-calibrate items in advance and store them in an *item bank* until they are needed on a new test form;
- *Test assembly:* Traditional paper-and-pencil tests can be designed to provide optimal measurement across a range of testee proficiencies; *computerized adaptive tests* can be designed to provide optimal measurement at or near the testee's true proficiency;
- *Differential item functioning:* Testing to see whether non-construct-related variation dominates or drives differences in item performance by different sociological groups;
- *Person-fit analysis:* Analogously to testing to see whether new items cohere with an existing item set, we can use the IRT model to test whether a person's response pattern is consistent with other peoples'; for example an unusual response pattern might have correct answers on hard items and incorrect answers on easy ones;
- And many more.

At the same time that IRT was finding widespread application in the engineering of large-scale assessments, as above, it was also being used in smaller-scale sociological and psychological assessments. In this context the nonparametric monotone homogeneity model of IRT was developed, to provide a framework in which scaling, scoring and person-fit questions might be addressed even though there was not enough data to adequately estimate a parametric IRT model. Later, in response to various anomalies that were uncovered in fitting unidimensional monotone IRT models to large-scale testing data, other forms of nonparametric IRT were developed, focusing, for example, on local dependence and accurate nonparametric estimation of IRFs.

In addition, IRT has been expanded in various ways to better account for the response process and data collection process. The LLTM, MLTM, NIDA/DINA models, and their generalizations, are all attempts to capture and measure finer-grained cognitive aspects of examinee performance. These models have in common that they are trying to stretch the IRT framework to accommodate a more modern and more detailed psychological view of the response process.

On the other hand, the Hierarchical Rater Model, as well as various testlet models, are designed to capture and correct for violations of LI due to the way the test was scored, or the way it was designed (e.g. several questions based on the same short reading). At the same time, demographic and other covariates are now routinely incorporated into IRT models for survey data, such as the National Assessment of Educational Progress in the United States (e.g. Allen, Donoghue and Schoeps, 2001, chap. 12), or PISA in Europe (e.g. OECD, 2005, chap. 9), to improve estimates of mean proficiencies in various demographic groups.

Thus, IRT and its antecedents have evolved, from an initial conceptual model that was

useful for defining and talking about what good measurement was, to a highly successful set of tools for engineering standardized testing, to a core component in a toolbox for rich statistical modeling of response and data collection processes in item-level discrete, direct response data. An initial disadvantage of this evolution is that connections with simple measurement criteria may be lost. However, the new IRT-based modeling toolbox is very flexible and can incorporate not only aspects of the data collection process (increasing its applicability generally), but also aspects of modern, detailed cognitive theory of task and test performance (increasing face validity and focusing inference on psychologically relevant constructs).

References

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, **7**, 255–278.

Adams, R.J., Wilson, M. & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, **21**, 1–23.

Allen, N.L., Donoghue, J.R., & Schoeps, T.L. (2001). *The NAEP 1998 Technical Report.* Washington, DC: National Center for Educational Statistics, U.S. Department of Education. Downloaded Sept. 29, 2005, from `http://nces.ed.gov/nationsreportcard/pubs/main1998/2001509.asp`.

Andersen, E.B. (1980). *Discrete statistical models with social science appplications.* Amsterdam: North-Holland.

Andrich, D. & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, **17**, 253–276.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika*, **66**, 471–488.

Binet, A., & Simon, Th. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, **11**, 191–244.

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.395–479). Reading: Addison-Wesley.

Bouwmeester, S., Sijtsma, K., & Vermunt, J.K. (2004). Latent class regression analysis for describing cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, **1**, 67–86.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, **64**, 153–168.

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

Coombs, C.H. (1964). *A theory of data.* Ann Arbor, MI: Mathesis Press.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297–334.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Davison, M. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika*, **42**, 523–548.

De Boeck, P. & Wilson, M. (2004; Eds.). *Explanatory Item Response Models: A Generalized*

Linear and Nonlinear Approach. New York: Springer Verlag.

De la Torre, J. & Douglas, J. (2004). Model evaluation and multiple strategies in cognitive diagnosis: an analysis of fraction subtraction data. In review.

DiBello, L.V., Stout, W.F., & Roussos, L.A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds). (1995). *Cognitively diagnostic assessment* (pp.361–389). Hillsdale, NJ: Erlbaum.

Douglas, J.A. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, **62**, 7–28.

Embretson, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, **56**, 495–515.

Embretson, S.E. (1997). Multicomponent response models. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.305–321). New York: Springer-Verlag.

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research*, **39**, 1-35.

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, **37**, 359–74.

Fischer, G.H., & Molenaar, I.W. (1995; Eds.). *Rasch models. Foundations, recent developments, and applications.* New York: Springer-Verlag.

Fraser, C., & McDonald, R.P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, **23**, 267–269.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian Data Analysis.* New York: Chapman-Hall/CRC.

Gibbons, R.D., & Hedeker, D.R. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, **53**, 1527–1537.

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer, & I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications* (pp.69–95). New York: Springer Verlag.

Guilford, J.P. (1936). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests.* Hillsdale, NJ: Erlbaum (reprinted in 1987).

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, **9**, 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction* (pp.60–90). Princeton, NJ: Princeton University Press.

Haberman, S.J. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics*, **5**, 815–841.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, **26**, 301–321.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, **62**, 331–347.

Hemker, B.T., Van der Ark, L.A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, **66**, 487–506.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 577–601.

Holland, P.W., and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, **14**, 1523–1543.

Jannarone, R.J. (1997). Models for locally dependent responses: conjunctive item response theory. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response*

theory (pp.465–479). New York: Springer-Verlag.

Janssen, R., & De Boeck, P. (1997). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement*, **21**, 37–50.

Johnson, M.S. (2005). Nonparametric estimation of item and respondent locations from unfolding-type items. *Psychometrika*, in press.

Johnson, M.S. & Junker, B.W. (2003). Using data augmentation and Markov chain Monte Carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics*, **28**, 195–230.

Junker, B.W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika*, **56**, 255–278.

Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics*, **21**, 1359–1378.

Junker, B.W. (1996). *Examining monotonicity in polytomous item response data*. Paper presented at the Annual Meeting of the Psychometric Society, June 27–30 1996, Banff, Alberta, Canada.

Junker, B.W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Prepared for the Committee on the Foundations of Assessment, National Research Council, Washington DC, November 30, 1999. Downloaded Sept. 26, 2005, from `http://www.stat.cmu.edu/~brian/nrc/cfa/`.

Junker, B.W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M.A.J. van Duijn, & T.A.B. Snijders (Eds.), *Essays on item response theory* (pp.247–276). New York: Springer-Verlag.

Junker, B.W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, **24**, 65–81.

Junker, B.W. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, **25**, 258–272.

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, **28**, 110–125.

Kelderman, H. & Rijkes, C.P.M. (1994). Loglinear multidimensional IRT models for polytomously scored items. *Psychometrika*, **59**, 149–176.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen (Eds.), *Measurement and prediction* (pp.362–412). Princeton: Princeton University Press.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. *Psychological Bulletin*, **45**, 507–530.

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

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

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

Mariano, L.T. & Junker, B.W. (2005). Covariates of the rating process in hierarchical models for multiple ratings of test items. Under review.

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, **60**, 523–546.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, **64**, 187–212.

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

Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, **19**, 91–100.

Meredith, W. (1965). Some results based on a general stochastic model for mental tests. *Psy-*

chometrika, **30**, 419–440.

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Erlbaum.

Mokken, R.J. (1971). *A theory and procedure of scale analysis.* De Gruyter, Berlin.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. In W.J. van der Linden & R.K. Hambleton, R.K. (Eds.), *Handbook of modern item response theory* (pp.351–367). New York: Springer.

Molenaar, I.W., & Sijtsma, K. (2000). *MSP5 for Windows. User's manual.* iecProGAMMA, Groningen, The Netherlands.

Muraki, E., & Carlson, J.E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, **19**, 73–90.

Neyman, J., & Scott, E.L. (1948). Consistent estimation from partially consistent observations. *Econometrica*, **16**, 1–32.

OECD (Organization for Economic Co-operation and Development (2005). PISA 2003 Technical Report. Paris: OECD. Downloaded Sept. 29, 2005, from `http://213.253.134.29/oecd/pdfs/browseit/9805011E.PDF`.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*. **27**, 341–384.

Post, W.J. (1992). *Nonparametric unfolding models: A latent structure approach.* DSWO Press, Leiden, The Netherlands.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, **56**, 611–630.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, **9**, 401–412.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.271–286). New York: Springer-Verlag.

Roberts, J.S., Donoghue, J.R., & Laughlin, J.E. (2000). A general model for unfolding unidimensional polytomous responses using item response theory. *Applied Psychological Measurement*, **24**, 3–32.

Rossi, N., Wang, X. and Ramsay, J.O. (2002) Nonparametric item response function estimates with the EM algorithm. *Journal of the Behavioral and Educational Sciences*, **27**, 291–317.

Rijkes, C.P.M. (1996). *Testing hypotheses on cognitive processes using IRT models.* Unpublished PhD thesis, Universiteit Twente, the Netherlands.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, No.17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No.18.*

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp.85–100). New York: Springer-Verlag.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **19**, 476–506.

Sijtsma, K., and Molenaar, I.W. (2002). *Introduction to nonparametric item response theory.* Sage, Thousand Oaks CA.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, **3**, 271–195.

Stegelmann, W. (1983). Expanding the Rasch model to a general model having more than one dimension. *Psychometrika*, **48**, 259–267.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, **52**, 589–617.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, **55**, 293–325.

Stout, W.F. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, **67**, 485–518.

Stout, W.F., Habing, B., Douglas, J., Kim, H., Roussos, L., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, **20**, 331–354.

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese*, **48**, 191–199.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman, and R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp.327–359). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, **49**, 501–519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, **51**, 567–577.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, **34**, 273–286.

Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika*, **11**, 1–13.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, **43**, 39–55.

Van der Ark, L.A. (2005). Practical consequences of stochastic ordering of the latent trait under various polytomous IRT models. *Psychometrika*, **70**, 283–304.

Van der Ark, L.A., Hemker, B.T., & Sijtsma, K. (2002). Hierarchically related nonparametric IRT models, and practical data analysis methods. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp.40–62). Manwah, NJ: Erlbaum.

Van der Linden, W.J. (2005). *Linear models for optimal test design.* New York: Springer.

Van der Linden, W.J., & Hambleton, R.K. (1997; Eds.). *Handbook of modern item response theory.* New York: Springer-Verlag.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory.* Unpublished doctoral dissertation, University of Amsterdam, The Netherlands.

Verhelst, N.D., & Mislevy, R.J. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, **55**, 195–215.

Verhelst, N.D. & Verstralen, H.H.F.M. (1993). A stochastic unfolding model derived from the partial credit model. *Kwantitative Methoden*, **42**, 73–92.

Walker, D.A. (1931). Answer pattern and score scatter in tests and examinations. *British Journal of Psychology*, **22**, 73–86.

Weaver, R.L., & Junker, B.W. (2003). *Model specification for cognitive assessment of proportional reasoning.* Technical Report #777, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA. Downloaded Sept. 28, 2005, from `http://www.stat.cmu.edu/tr/tr777/tr777.pdf`.

Wilson, D., Wood, R.L., & Gibbons, R. (1983). TESTFACT: Test scoring and item factor analysis. [Computer software]. Chicago: Scientific Software Inc.

Zhang, J., & Stout, W.F. (1999a). Conditional covariance structure of generalized compensatory

multidimensional items. *Psychometrika*, **64**, 129–152.
Zhang, J., & Stout, W.F. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, **64**, 213–249.

(Received October 4 2005, Revised January 26 2006)