1

Item Response Theory: Brief History, Common Models, and Extensions

Wim J. van der Linden and Ronald K. Hambleton

Long experience with measurement instruments such as thermometers, yardsticks, and speedometers may have left the impression that measurement instruments are physical devices providing measurements that can be read directly off a numerical scale. This impression is certainly not valid for educational and psychological tests. A useful way to view a test is as a series of small experiments in which the tester records a vector of responses by the testee. These responses are not direct measurements, but provide the data from which measurements can be inferred.

Just as in any experiment, the methodological problem of controlling responses for experimental error holds for tests. Experimental error with tests arises from the fact that not only the independent variables but also other variables operational in the experiment, generally known as background, nuisance, or error variables, may have an impact on the responses concerned. Unless adequate control is provided for such variables, valid inferences cannot be made from the experiment.

Literature on experimental design has shown three basic ways of coping with experimental error (Cox, 1958):

(1) matching or standardization;

(2) randomization; and

(3) statistical adjustment.

If conditions in an experiment are matched, subjects operate under the same levels of error or nuisance variables, and effects of these variables cannot explain any differences in experimental outcomes. Although matching is a powerful technique, it does have the disadvantage of restricted generalizability. Experimental results hold only conditionally on the levels of the matching variables that were in force.

Randomization is based on the principle that if error variables cannot be manipulated to create the same effects, random selection of conditions guarantees that these effects can be expected to be the same on average.

W. J. van der Linden et al. (eds.), *Handbook of Modern Item Response Theory*

Thus, these effects can not have a systematic influence on the responses of the subjects.

Statistical adjustment, on the other hand, is a technique for *post hoc* control, that is, it does not assume any intervention from the experimenter. The technique can be used when matching or randomization is impossible, but the technique does require that the levels of the relevant error variables have been measured during the experiment. The measurements are used to adjust the observed values of the dependent variables for the effects of the error variables.

Adjustment procedures are model-based in the sense that quantitative theory is needed to provide the mathematical equations relating error variables to the dependent variables. In a mature science such as physics, substantive theory may be available to provide these equations. For example, if temperature or air pressure are known to be disturbing variables that vary across experimental conditions and well-confirmed theory is available to predict how these variables affect the dependent variable, then experimental findings can be adjusted to common levels of temperature or pressure afterwards. If no substantive theory is available, however, it may be possible to postulate plausible equations and to estimate their parameters from data on the error and dependent variables. If the estimated equations fit the data, they can be used to provide the adjustment. This direction for analysis is popular in the behavioral sciences, in particular in the form of analysis of covariance (ANCOVA), which postulates linear (regression) equations between error and dependent variables.

Classical Test Theory

Classical test theory fits in with the tradition of experimental control through matching and randomization (van der Linden, 1986). The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in the ability (i.e., true score) of interest. All other potential sources of variation existing in the testing materials, external conditions, or internal to the examinees are assumed either to be constant through rigorous standardization or to have an effect that is nonsystematic or "random by nature."

The classical test theory model, which was put forward by Spearman (1904) but received its final axiomatic form in Novick (1966), decomposes the observed test score into a true score and an error score. Let X_{jg} be the observed-score variable for examinee j on test g, which is assumed to be random across replications of the experiment. The classical model for a fixed examinee j postulates that

$$X_{jg} = \tau_{jg} + E_{jg}, \tag{1}$$

where τ_{jg} is the true score defined as the expected observed score, $\mathcal{E}_{jg}X_{jg}$, and $E_{jg} \equiv X_{jg} - \tau_{jg}$ is the error score. Note that, owing to the definition of the true and error scores, the model does not impose any constraints on X_{jg} and is thus always "true." Also, observe that τ_{jg} is *defined* to cover all systematic information in the experiment, and that it is only a parameter of interest if experimental control of all error variables has been successful. Finally, it is worthwhile to realize that, as in any application of the matching principle, τ_{jg} has only meaning conditionally on the chosen levels of the standardized error variables. The true score is fully determined by the test as designed—not by some Platonic state inside the examinee that exists independent of the test (Lord and Novick, 1968, Sec. 2.8).

If random selection of examinees from a population is a valid assumption, the true score parameter, τ_{jg}, has to be replaced by a random true score variable, T_{j*}. The model in Eq. (1) becomes:

$$X_{j*} = T_{j*} + E_{j*} \tag{2}$$

(Lord and Novick, 1968, Sec. 2.6). It is useful to note that Eq. (2) is also the linear equation of the one-way analysis of variance (ANOVA) model with a random factor. This equivalence points again to the fact that educational and psychological tests can be viewed as regular standardized experiments.

Test-Dependent True Scores

As mentioned earlier, a serious disadvantage of experiments in which matching is used for experimental control is reduction of external validity. Statistical inferences from the data produced by the experiment cannot be generalized beyond the standardized levels of its error or nuisance variables. The same principle holds for standardized tests. True scores on two different tests designed to measure the same ability variable, even if they involve the same standardization of external and internal conditions, are generally unequal. The reason is that each test entails its own set of items and that each item has different properties. From a measurement point of view, *such properties of items are nuisance or error variables that escape standardization.*

The practical consequences of this problem, which has been documented numerous times in the test-theoretic literature [for an early reference, see Loevinger (1947)], are formidable. For example, it is impossible to use different versions of a test in a longitudinal study or on separate administrations without confounding differences between test scores by differences between the properties of the tests.

Statistical Adjustment for Differences Between Test Items

Perhaps the best way to solve the problem of test-dependent scores is through the third method of experimental control listed previously—statistical adjustment. The method of statistical adjustment requires explicit parametrization of the ability of interest as well as the properties of the items, via a model that relates their values to response data collected through the test. If the model holds and the item parameters are known, the model adjusts the data for the properties of items in the test, and, therefore, can be used to produce ability measures that are free of the properties of the items in the test. This idea of statistical adjustment of ability measures for nuisance properties of the items is exactly analogous to the way analysis of covariance (ANCOVA) was introduced to parametrize and subsequently "remove" the effects of covariates in an ANOVA study.

Item Response Theory

Mathematical models to make statistical adjustments in test scores have been developed in item response theory (IRT). The well-known IRT models for dichotomous responses, for instance, adjust response data for such properties of test items as their difficulty, discriminating power, or liability to guessing. These models will be reviewed in this chapter.

The emphasis in this volume, however, it not on IRT models for handling dichotomously scored data, but is on the various extensions and refinements of these original models that have emerged since the publication of Lord and Novick's *Statistical Theories of Mental Test Scores* (1968). The new models have been designed for response data obtained under less rigorous standardization of conditions or admitting open-ended response formats, implying that test scores have to be adjusted for a larger array of nuisance variables. For example, if, in addition to the ability of interest, a nuisance ability has an effect on responses to test items, a second ability parameter may be added to the model to account for its effects. Several examples of such multidimensional IRT models are presented in this volume. Likewise, if responses are made to the individual answer choices of an item (e.g., multiple-choice item), they clearly depend on the properties of the answer choices, which have to be parametrized in addition to other properties of the item. A variety of models for this case, generally known as polytomous IRT models, is also presented in this volume.

The methodological principle underlying all of the models included in this volume is the simple principle of a separate parameter for each factor with a separate effect on the item responses. However, it belongs to the "art of modeling" to design a mathematical structure that reflects the interaction between these factors and at the same time makes the model statistically tractable. Before introducing the models in a formal way, a brief review of common dichotomous IRT models is given in this chapter. These were the first IRT models to be developed, and they have been widely researched and extensively used to solve many practical measurement problems (Hambleton and Swaminathan, 1985; Lord, 1980).

Review of Dichotomous IRT Models

As its name suggests, IRT models the test behavior not at the level of (arbitrarily defined) test scores, but at the item level. The first item format addressed in the history of IRT was the dichotomous format in which responses are scored either as correct or incorrect. If the response by examinee j to item i is denoted by a random variable U_{ij}, it is convenient to code the two possible scores as $U_{ij} = 1$ (correct) and $U_{ij} = 0$ (incorrect). To model the distribution of this variable, or, equivalently, the probability of a correct response, the ability of the examinee is presented by a parameter $\theta \in (-\infty, +\infty)$, and it is assumed in a two-parameter model that the properties of item i that have an effect on the probability of a success are its "difficulty" and "discriminating power," which are represented by parameters $b_i \in (-\infty, +\infty)$ and $a_i \in (0, +\infty)$, respectively. These parameters are simply symbols at this point; their meaning can only be established if the model is further specified.

Since the ability parameter is the *structural parameter* that is of interest and the item parameters are considered *nuisance parameters*, the probability of success on item i is usually presented as $P_i(\theta)$, that is, as a function of θ specific to item i. This function is known as the *item response function* (IRF). Previous names for the same function include *item characteristic curve* (ICC), introduced by Tucker (1946), and *trace line*, introduced by Lazarsfeld (1950). Owing to restrictions in the range of possible probability values, item response functions cannot be linear in θ. Obviously, the function has to be monotonically increasing in θ. The need for a monotonically increasing function with a lower and upper asymptote at 0 and 1, respectively, suggests a choice from the class of cumulative distributions functions (cdf's). This choice was the one immediately made when IRT originated.

Normal-Ogive Model

The first IRT model was the normal-ogive model, which postulated a normal cdf as a response function for the item:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/z} dz. \tag{3}$$

For a few possible values of the parameters b_i and a_i, Figure 1 displays the shape of a set of normal-ogive response functions. From straightforward analysis of Eq. (3) it is clear that, as demonstrated in Figure 1, the difficulty parameter b_i is the point on the ability scale where an examinee has a probability of success on the item of 0.50, whereas the value of a_i is proportional to the slope of the tangent to the response function at this point. Both formal properties of the model justify the interpretation suggested by the names of the parameters. If b_i increases in value, the response

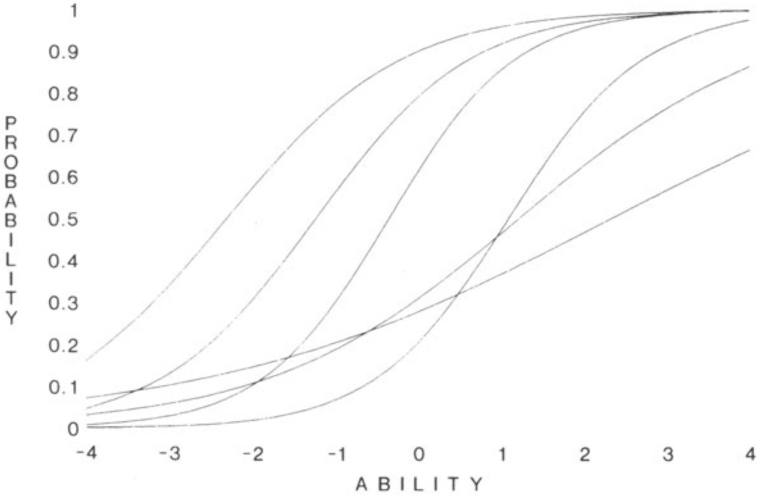

FIGURE 1. Two-parameter normal-ogive response functions.

function moves to the right and a higher ability is needed to produce the same probability of success on the item. Also, the larger the value of a_i, the better the item discriminates between the probabilities of success of examinees with abilities below and above $\theta = b_i$.

Although credit for the normal-ogive model is sometimes given to Lawley (1943), the same basic model had been studied earlier by Ferguson (1942), Mosier (1940, 1941), and Richardson (1936). In fact, the original idea for the model can be traced to Thurstone's use of the normal model in his discriminal dispersions theory of stimulus perception (Thurstone, 1927b). This impact of the psychophysical scaling literature on the early writings of these authors is obvious from their terminology. Researchers in psychophysics study the relation between the physical properties of stimuli and their perception by human subjects. Its main method is to present a stimulus of varying strength, for example, a tone of varying loudness, where the subject's task is to report whether or not he/she was able to detect the stimulus. Since the probability of detection is clearly an increasing function of the physical strength of the stimulus, the normal cdf with the parametrization in Eq. (3) was used as response function. However, in this model, θ stands for the *known* strength of a stimulus and the interest was only in the parameters a_i and b_i, the latter being known as the *limen* of the stimulus—a name also used for the difficulty of the item in the early days of IRT. Mosier (1940, 1941) was quite explicit in his description of the parallels between psychophysics and psychometric modeling. In his 1940 publication he provides a table with the different names used in the two disciplines, which in fact describe common quantities.

Early descriptions of the normal-ogive model as well as of the procedures for its statistical analysis were not always clear. For example, ability θ was not always considered as a latent variable with values to be estimated from the data but taken to be the observed test score. Equation (3) was used as a model for the (nonlinear) regression of item on test scores. Richardson, on the other hand, used the normal-ogive model to scale a dichotomous external success criterion on the observed-score scale of several test forms differing in difficulty. In the same spirit that originated with Binet and Simon (1905), and was reinforced by the widespread use of Thurstone's method of absolute scaling (Thurstone, 1925, 1927a), the ability variable was always assumed to be normally distributed in the population of examinees. The belief in normality was so strong that even if an observed ability distribution did not follow a normal distribution, it was normalized to find the "true" scale on which the model in Eq. (3) was assumed to hold. Also, in accordance with prevailing practice, the ability scale was divided into a set of—typically seven—intervals defined by equal standard-deviation units. The midpoints of the intervals were the discrete ability scores actually used to fit a normal-ogive model.

The first coherent treatment of the normal-ogive model that did not suffer from the above idiosyncrasies was given in Lord (1952). His treatment also included psychometric theory for the bivariate distribution of item and ability scores, the limiting frequency distributions of observed scores on large tests, and the bivariate distribution of observed scores on two tests measuring the same ability variable. All these distributions were derived with the help of the normal-ogive model.

Parameter Estimation. In the 1940s and 1950s, computers were not available and parameter estimation was a laborious job. The main estimation method was borrowed from psychophysics and known as the *constant process.* The method consisted of fitting a weighted regression line through the data points and the empirical probits. [The latter were defined as the inverse transformation of Eq. (3) for empirical proportions of successes on the item.] The weights used in the regression analysis were known as the Müller–Urban weights. Lawley (1943) derived maximum-likelihood (ML) estimators for the item parameters in the normal-ogive model and showed that these were identical to the constant-process estimators upon substitution of empirical probits in the Müller–Urban weights.

Lord and Novick (1968, Secs. 16.8–16.10) presented the following set of equations, derived under the assumption of a normal ability distribution, that relate the parameters b_i and a_i to the classical item-π value and biserial item-test correlation, ρ_i:

$$a_i = \frac{\rho_i}{\sqrt{1 - \rho_1^2}}, \tag{4}$$

$$b_i = \frac{-\gamma_i}{\rho_i}, \tag{5}$$

where γ_i is defined by

$$\gamma_i = \Phi^{-1}(\pi_i) \tag{6}$$

and $\Phi(\cdot)$ is the standard normal distribution function. Although the equations were based on the assumption of the normal-ogive model, plug-in estimates based on these equations have long served as heuristic estimates or as starting values for the maximum likelihood (ML) estimators of the parameters in the logistic model (Urry, 1974).

Goodness of Fit. Formal goodness-of-fit tests were never developed for the normal-ogive model. In the first analyses published, the basic method of checking the model was graphical inspection of plots of predicted and empirical item-test regression functions [see, for example, Ferguson (1942) and Richardson (1936)]. Lord (1952) extended this method to plots of predicted and empirical test-score distributions and used a chi-square test to study the differences between expected and actual item performance at various intervals along the ability continuum.

Rasch or One-Parameter Logistic Model

Rasch began his work in educational and psychological measurement in the late 1940s. In the 1950s he developed two Poisson models for reading tests and a model for intelligence and achievement tests. The latter was called "a structural model for items in a test" by him, but is now generally known as the Rasch model. Formally, the model is a special case of the Birnbaum model to be discussed below. However, because it has unique properties among all known IRT models, it deserves a separate introduction. A full account of the three models is given in Rasch (1960).

Rasch's main motivation for his models was his desire to eliminate references to populations of examinees in analyses of tests (Rasch, 1960, Preface; Chap. 1). Test analysis would only be worthwhile if it were individual-centered, with separate parameters for the items and the examinees. To make his point, Rasch often referred to the work of Skinner, who was also known for his dislike of the use of population-based statistics and always experimented with individual cases. Rasch's point of view marked the transition from population-based classical test theory, with its emphasis on standardization and randomization, to IRT with its probabilistic modeling of the interaction between an individual item and an individual examinee. As will be shown, the existence of sufficient statistics for the item parameters in the Rasch model can be used statistically to adjust ability estimates for the presence of nuisance properties of the items in a special way.

Poisson Model. Only the model for misreadings will be discussed here. The model is based on the assumption of a Poisson distribution for the number of reading errors in a text. This assumption is justified if the reading process is stationary and does not change due to, for example, the examinee becoming tired or fluctuations in the difficulties of the words in the text.

If the text consists of T words and X is the random variable denoting the number of words misread, then the Poisson distribution assumes that

$$P(X = x \mid T) = e^{-\lambda} \frac{\lambda^x}{x!}, \tag{7}$$

where parameter λ is the expected number of misreadings. Equivalently, $\xi \equiv \lambda/T$ is the probability of misreading a single word sampled from the text.

The basic approach followed by Rasch was to further model the basic parameter ξ as a function of parameters describing the ability of the examinee and the difficulty of the text. If θ_j is taken to represent the reading ability of examinee j and δ_t represents the difficulty of text t, then ξ_{it} can be expected to be a parameter decreasing in θ_j and increasing in δ_t. The simple model proposed by Rasch was

$$\xi_{ij} = \frac{\delta_t}{\theta_j}. \tag{8}$$

If the model in Eqs. (7) and (8) is applied to a series of examinees $j = 1, \ldots, N$ reading a series of texts $t = 1, \ldots, T$, it holds that the sum of reading errors in text t across examinees, $X_{t.}$, is a sufficient statistic for δ_t. It is a well-known statistical result that the distribution of X_{tj} given $X_{t.} = x_{t.}$ follows a binomial distribution with a success parameter independent of δ_t. Removing the effect of text difficulty from the data when estimating the ability of an examinee is realized when the inference is based on the conditional likelihood function associated with this binomial.

Rasch Model. The main model for a test with items $i = 1, \ldots, n$ proposed by Rasch departs from the parameter structure defined in Eq. (8). Substituting item difficulty δ_i for text difficulty δ_t, the question can be asked about the simplest model with the parameter structure in Eq. (8). The answer proposed by Rasch was the following transformation:

$$\begin{aligned} P_i(U_{ij} = 1 \mid \theta) &= \frac{\frac{\theta_j}{\delta_i}}{1 + \frac{\theta_j}{\delta_i}} \\ &= \frac{\theta_j}{\theta_j + \delta_i}. \end{aligned} \tag{9}$$

The model, presented in this form, has the simple interpretation of the probability of success being equal to the value of the person parameter relative to the value of the item parameter. Taking the parameters on a logarithmic scale (but maintaining the notation), the model can easily be

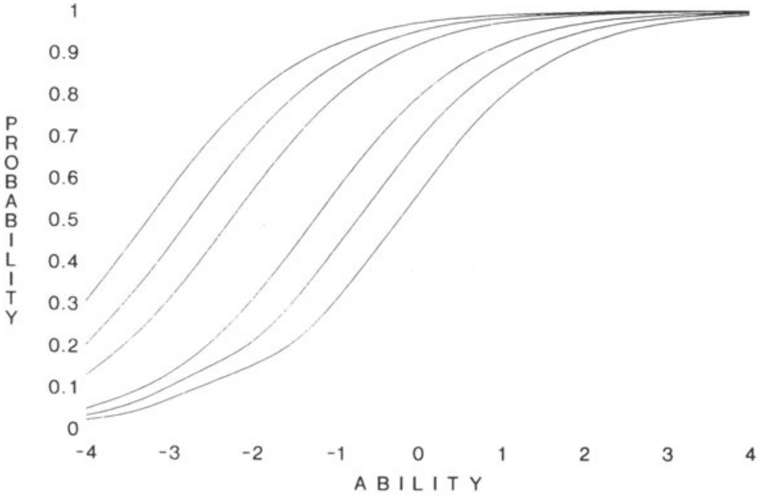

FIGURE 2. One-parameter logistic response functions.

shown to be equivalent to the following form, known as the one-parameter logistic (or 1-PL) model:

$$P(U_{ij}=1 \mid \theta) = \frac{1}{1+\exp\{-(\theta-\delta_i)\}}. \tag{10}$$

This representation can be used to show a fundamental property of the Rasch model relative to the two-parameter normal-ogive model, namely, that its IRF's do not intersect. Figure 2 illustrates this property graphically for a few IRF's.

As demonstrated in the next section, the model uses the same form of "item-free" estimation of examinee ability as the Poisson model described above.

Parameter Estimation. Using the representation in Eq. (9), the joint likelihood function associated with the case of N examinees having responded to n items is equal to

$$\begin{aligned} L(\theta,\delta;u) &= \prod_i \prod_j \frac{(\theta_j/\delta_i)^{u_{ij}}}{1+\theta_j/\delta_i} \\ &= \frac{\prod_i \delta_i^{-u_{i.}} \prod_j \theta_j^{u_{.j}}}{\prod_i \prod_j (1+\theta_j/\delta_i)}, \end{aligned} \tag{11}$$

where $\theta \equiv (\theta_1,\ldots,\theta_N)$, $\delta \equiv (\delta_1,\ldots,\delta_n)$, $u \equiv (u_{ij})$, and $u_{i.}$ and $u_{.j}$ are the marginal sums of the data matrix. Since the marginal sums can be shown to be sufficient statistics for the difficulty and item parameters, respectively, it holds for the likelihood function associated with the conditional probability of (U_{ij}) given $U_i = u_i$ that

$$L(\theta; u \mid u_{i.}) = \frac{\prod_j \theta^{u_{j.}}}{\prod_i \gamma_{u_{i.}}}, \tag{12}$$

where the functions $\gamma_{u_{i.}}$ are combinatorial functions of the ability parameters known as elementary symmetric functions (Fischer, 1974, Sec. 13.5).

Three comments on this result should be made. First, the conditional likelihood function contains only the ability parameters. Maximum-likelihood estimates of the ability parameters can thus be obtained from (conditional) likelihood functions free of any item parameter. It is in this sense that sufficient statistics for the item parameters can be used to adjust the data for the difficulty of the items. (Note, however, that these CML estimators are not independent of the item parameters. They have small-sample first moments that may depend on the item parameters, whereas their second moments always depend on these parameters.) The result in Eq. (12) holds for any model in the exponential family, which includes the Rasch model (Lehmann, 1959, Theor. 1.2.1).

Second, the same procedure can be followed to obtain CML estimates for the item parameters. In practice, however, only the item parameters are estimated by CML estimation. These parameters are then considered as known and the examinee parameters are estimated under this assumption by regular maximum likelihood methods.

Third, until recently, the numerical problems involved in dealing with the elementary symmetric functions in the CML equations seemed insurmountable, and applications to tests with more than 60 items were not feasible. The situation has changed favorably since new algorithms with improved efficiency have become available that permit CML estimation of hundreds of parameters (Liou, 1994; Verhelst et al., 1984).

As an alternative to CML estimation, estimates of the item parameters can be found through maximization of the marginal likelihood function, obtained from Eq. (11) by integrating over a common density for the θ's, which is typically taken to be a normal density with parameters μ and σ:

$$L(\delta; u, \mu, \sigma) = \int_{\infty}^{-\infty} L(\delta, \theta; u) f(\theta; \mu, \sigma) d\theta. \tag{13}$$

An efficient way to calculate these marginal maximum-likelihood (MML) estimates is to use an EM algorithm (Thissen, 1982). Marginalization as in Eq. (13) is another way to eliminate the impact of the ability parameters when estimating item parameters.

A variety of other estimation procedures have been studied, including (hierarchical) Bayesian methods (Swaminathan and Gifford, 1982), semi-parametric methods in which the ability density in Eq. (13) is estimated from the data (de Leeuw and Verhelst, 1986; Engelen, 1989), an iterative least-squares method (Verhelst and Molenaar, 1988), and a minimum chi-square method (Fischer, 1974, Sec. 14.8).

Goodness of Fit

Rasch was not particularly interested in formal tests of the goodness of fit of his model. In his 1960 book, in a fashion which was similar to the early work with the normal-ogive model, he only diagnosed the fit of his model using plots of residuals.

The first statistical test for the goodness of fit of the Rasch model was Andersen's (1973) likelihood-ratio test. The test was designed to check whether item parameter estimates in different score groups are equal up to the sampling errors of the estimates. Several other statistical tests have followed. Molenaar (1983) offered a statistic sensitive to differences in discriminating power between items. Van den Wollenberg (1982) introduced the test statistics Q_1 and Q_2, the second of which is sensitive to violations of the assumption of unidimensionality. Large families of possible tests, with power against specific violations of the model, are given in Glas (1988, 1989) and Kelderman (1984).

A controversial aspect of analyzing test items for model fit has always been whether the case of a misfitting item should lead to removal of the item from the test or to a relaxation of the model. In principle, both strategies are possible. Test practice does have a long tradition of item analysis in which items not meeting certain conventional values for the classical difficulty and discrimination parameters are removed from the pool or the test. On the other hand, a relaxation of the Rasch model is available in the form of the two- and three-parameter logistic models, which may better fit problematic items (but need not necessarily do so). The position taken here is that no general recommendation can be made with respect to this choice between a more stringent model with excellent statistical tractability and a more flexible model likely to fit a larger collection of items. Additional factors such as (1) the nature of the misfit, (2) the availability of substitute items, (3) the amount of time available for rewriting items, (4) the availability of a sufficiently large sample to properly estimate item parameters for more general models, and—probably most important—(5) the goal of the testing procedure play a significant role in the handling of items that are not fit by the Rasch model. For example, different decisions about the handling of model misfit are likely to be made in a long-term project developing a psychological test versus an educational assessment project with an item pool that has to cover comprehensively the content of a textbook used by the schools.

Birnbaum's Two- and Three-Parameter Logistic Models

Birnbaum worked on his contributions to IRT in the late 1950s but his work became widely known through the chapters he contributed to Lord and Novick (1968, Chaps. 17–20). Unlike Rasch, Birnbaum's work was not motivated by a desire to develop a different kind of test theory. As a statistician, his main aim was to make the work begun by Lord (1952) on the normal-ogive model statistically feasible. In particular, he provided the statistical theory for ability and item parameter estimation, applied statistical information measures to ability estimation and hypothesis testing with respect to ability values, and proposed a rational approach to test construction.

Birnbaum's main contribution, however, was his suggestion to replace the normal-ogive model in Eq. (3) by the logistic model:

$$P_i(\theta) = \frac{1}{1 + \exp\{-a_i(\theta - b_i)\}}\,. \tag{14}$$

The motivation for this substitution was the result of Haley (1952) that for a logistic cdf with scale factor 1.7, $L(1.7x)$, and a normal cdf, $N(x)$:

$$|N(x) - L(1.7x)| < 0.01 \qquad \text{for } x \in (-\infty, \infty)$$

[for a slight improvement on this result, see Molenaar (1974)]. At the same time, the logistic function is much more tractable than the normal-ogive function, while the parameters b_i and a_i retain their graphical interpretation shown in Figure 1.

Birnbaum also proposed a third parameter for inclusion in the model to account for the nonzero performance of low-ability examinees on multiple-choice items. This nonzero performance is due to the probability of guessing correct answers to multiple-choice items. The model then takes the form

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp\{-a_i(\theta - b_i)\}}\,. \tag{15}$$

Equation (15) follows immediately from the assumption that the examinee either knows the correct response with a probability described by Eq. (14) or guesses with a probability of success equal to the value of c_i. From Figure 3, it is clear that the parameter c_i is the height of the lower asymptote of the response function. In spite of the fact that Eq. (15) no longer defines a logistic function, the model is known as the three-parameter logistic (or 3-PL) model, while the model in Eq. (14) is called the two-parameter logistic (or 2-PL) model. The c-parameter is sometimes referred to as the "guessing parameter," since its function is to account for the test item performance of low-ability examinees.

One of Birnbaum's other contributions to psychometric theory was his introduction of Fisher's measure to describe the information structure of a

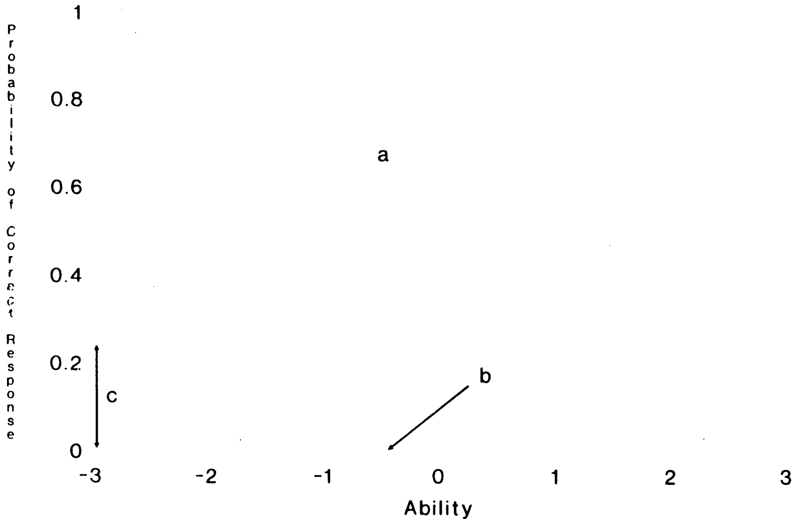

FIGURE 3. Three-parameter logistic response function.

test. For a dichotomous IRT model, Fisher's measure for the information on the unknown ability θ in a test of n items can be written as

$$\begin{aligned} I(\theta) &= \sum_{i=1}^{n} I_i(\theta) \\ &= \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]}, \end{aligned} \tag{16}$$

where $I_i(\theta)$ is the information on θ in the response on item i and $P_i'(\theta) \equiv \frac{\partial}{\partial\theta} P_i(\theta)$. One of his suggestions for the use of these information measures was to influence test construction by first setting up a target for the information function of the test and then assembling the test to meet this target using the additivity in Eq. (16).

Birnbaum also derived weights for optimally scoring a test in a two-point classification problem and compared the efficiency of several other scoring formulas for the ability parameter in the model. His strongest statistical contribution no doubt was his application of the theory of maximum likelihood to the estimation of both the ability and the item parameters in the logistic models, which will now be addressed.

Parameters Estimation. Continuing the same notation as before, the likelihood function for the case of N examinees and n items associated with the 2-PL model can be written as

$$L(\theta, a, b; u) = \prod_i \prod_j P_i(\theta_j; a_i, b_i)^{u_{ij}} [1 - P_i(\theta_j; a_i, b_i)]^{1-u_{ij}}. \tag{17}$$

Maximizing the logarithm of the likelihood function results in the following set of estimation equations:

$$\begin{aligned}
&\sum_i a_i(u_{ij} - P_i(\theta_j; a_i, b_i)) = 0, \qquad j = 1, \ldots, N;\\
&\sum_j a_i(u_{ij} - P_i(\theta_j; a_i, b_i)) = 0, \qquad i = 1, \ldots, n; \qquad (18)\\
&\sum_j (u_{ij} - P_i(\theta_j; a_i, b_i))(\theta_j - b_i) = 0, \qquad i = 1, \ldots, n.
\end{aligned}$$

The technique suggested by Birnbaum was jointly to solve the equations for the values of the unknown parameters in an iterative scheme, which starts with initial values for the ability parameters, solves the equations for the item parameters, fixes the item parameters, and solves the equations for improved estimates of the values of the ability parameters, etc. The solutions obtained by this procedure are known as the joint maximum likelihood (JML) estimates, although the name alternating maximum likelihood (AML) would seem to offer a better description of the procedure.

Problems of convergence and a somewhat unknown statistical status of the JML estimators have led to the popularity of MML estimation for the 2-PL and 3-PL models. The method was introduced by Bock and Lieberman (1970) for the normal-ogive model but immediately adopted as a routine for the logistic model. For a density function $f(\theta)$ describing the ability distribution, the marginal probability of obtaining a response vector $u \equiv (u_1, \ldots, u_n)$ on the items in the test is equal to

$$P(u \mid a, b) = \int_{-\infty}^{\infty} \prod_i P_i(\theta; a_i, b_i)^{u_i}[1 - P_i(\theta; a_i, b_i)]^{1-u_i} f(\theta) d\theta. \qquad (19)$$

The marginal likelihood function associated with the full data matrix $u \equiv (u_{ij})$ is given by the multinomial kernel

$$L(a, b; u) = \prod_{u=1}^{2^n} \pi_u^{r_u}, \qquad (20)$$

where π_u is the probability of the response pattern in Eq. (19), which has frequency r_u in the data matrix. Owing to the large number of possible response vectors, the Bock and Lieberman method was slow and not practical for more than about 10 items. However, Bock and Aitkin (1981) introduce a version of the method in which an EM algorithm is implemented. This version is able to handle realistic numbers of items and has become the popular approach for the estimation of item parameters in the 2-PL and 3-PL model.

It is clear from the likelihood function in Eq. (20), that the ability parameters have been removed. Therefore, MML estimators of the item parameters are estimators that share such properties of regular ML estimators as consistency and asymptotic efficiency (Andersen, 1980, Chap. 2). The assumption of a density for the ability distribution needed to arrive at Eq. (20), for which generally the standard normal density is chosen, is mild. For an extensive description of an application of MML estimation with the EM algorithm to actual test data, see Tsutakawa (1984). For more details on ML estimation in these models, see Baker (1992) and Hambleton and Swaminathan (1985, Chaps. 5 and 7).

A variety of Bayesian approaches to parameter estimation in the 2-PL and 3-PL model have been studied. Relevant references are Mislevy (1986), Swaminathan and Gifford (1985, 1986), and Tsutakawa and Lin (1986).

Ability estimation with known item parameters may produce likelihood functions with multiple modes (Samejima, 1973; Yen et al., 1991). The 3-PL model also suffers from a problem that is known in the context of linear models as multicollinearity. From Fig. 3, it is clear that, for an interval of ability values about the location of the IRF, small changes in the value of the lower asymptote can be compensated for by small changes in the slope of the curve. Therefore, estimates of a_i and c_i are sensitive to minor fluctuations in the responses used to produce these estimates. Unless huge samples of examples or tight priors around the true parameter values are used, the estimates unstable.

Goodness of Fit. Whereas with the Rasch model, a number of sound statistical tests have been produced for addressing fit, the situation for the two- and three-parameter models is quite different. Well-established statistical tests do not exist, and even if they did, questions about the utility of statistical tests in assessing model fit can be raised, especially with large samples. Since no model is likely to fit perfectly a set of test items, given sufficient amounts of data, the assumption of model fit or adequacy, whatever the model, is likely to be rejected. Clearly then, statistical tests should not be used solely to determine the adequacy of model fit. It is quite possible that the departures between test data and predictions from the best fitting model are of no practical consequence but that a statistical test of model fit would reject the null hypothesis of model adequacy. It has been said that failure to reject an IRT model is simply a sign that sample size was too small (McDonald, 1989).

The approach to the assessment of fit with the two- and three-parameter models often involves the collection of a wide variety of evidence about model fit, including statistical tests, and then making an informed judgment about model fit and usefulness of a model with a particular set of data (Hambleton, 1989). This approach often includes fitting multiple models to test data and comparing the results and the practical consequences of any differences. It is common to carry out three types of studies: (1) checks on model assumptions, (2) checks on model parameter invariance, and (3) checks on predictions from the model. The basic idea is to become familiar with the model (and other possible models), the test data, model fit, and the consequences of model misfit, prior to making any decision about the utility of a particular model.

Checks on model assumptions will almost certainly involve investigations of the extent to which the unidimensionality assumption is valid. Common checks might include analysis of (1) the eigenvalues from the matrix of interitem correlations, (2) residuals after fitting a one-factor linear or nonlinear factor analytic model, or (3) IRT item statistics obtained in the total test and in a subset of items from the test that might potentially measure a second ability. For a useful review of approaches to the study of unidimensionality, Hattie (1985) is an excellent resource. Other checks on the test data might consist of a review of item biserial correlations (if variation is substantial, the usefulness of the two- and three-parameter models is increased), test format and difficulty (with multiple-choice items and especially if the items are difficult, a guessing parameter in the model could be helpful), and test speededness (test speededness is a threat to the assumption of unidimensionality).

Checks on model parameter invariance can be carried out. Here, the model of interest is assumed to be true and model parameter estimates are obtained. If the model fit is acceptable, examinee ability estimates ought to be the same (apart from measurement error) from different samples of items within the test (e.g., easy versus hard). Item parameter estimates ought to be about the same (except for sampling errors) from different samples of examinees from the population of examinees for whom the test is intended (e.g., males and females). Checks on the presence of ability and item parameter invariance provide valuable information about model fit (see Hambleton, Swaminathan and Rogers, 1991).

Predictions from the model, assuming it to be true or correct, can also be checked. Analysis of residuals or standardized residuals for items as well as persons is common via the study of residual plots (Hambleton, 1989). Statistical tests of departures in the data from model predictions can be carried out too [e.g., Yen (1981)]. Comparisons of residuals from fitting multiple models is especially helpful in determining the adequacy of model fit. Other studies that might be carried out include the comparison of predicted score distributions with the actual distributions (from several models) and the consequences of model fit (such as the ranking of examinee abilities under different models).

Historical Remark

The above review of IRT followed the history of models for dichotomous responses as they were developed in mental test theory. A potential danger of such treatments is to ignore parallel developments in other disciplines. It was mentioned earlier in the chapter that the first attempts to introduce the normal-ogive model in test theory were motivated by the successful application of the same model in psychophysical scaling. A parallel development is found in the work on models for dose-response curves in bioassay. In this field, the use of the normal-ogive model was popular until the early 1940s when it was replaced by the logistic model. An advocate of this change was Berkson (1944, 1951) who later worked on the statistical refinements of the use of the logistic model (Berkson 1953, 1955). Probably Berkson's work motivated Birnbaum to apply the same model in test theory. Baker (1992), in his book on parameter estimation in IRT, carefully spelled out the various historic links between methods used in bioassay and test theory.

Another main stimulus to work in IRT was Lazarsfeld's work on latent structure analysis. Although the major portion of Lazarsfeld's work was devoted to models for a finite number of latent classes of examinees with different probabilities of responding to items, his earlier work also included models for continuous response functions such as the polynomial, linear, and latent-distance models [for an overview, see Lazarsfeld and Henry (1968)]. Lazarsfeld's most important impact on test theory, however, was methodological. He introduced the terminology to distinguish between latent parameters of interest and manifest data and convinced test theorists that probabilistic models were needed to account for the relation between the two quantities. Lazarsfeld's latent class models have had a larger impact on sociology than on education and psychology. However, some of the modern IRT models in this volume have built the presence of latent classes into IRT models for continuous ability variables, for example, to represent differences between examinees in solution strategies used to solve test items (Kelderman, this volume; Rost, this volume).

Finally, an important link exists between the normal-ogive model and work on factor analysis of dichotomous variables. Both the normal-ogive and the linear factor model dealt with latent quantities, and the question about the relationship between the models has been of concern to several researchers. Lord and Novick (1968, Sec. 16.8) proved that a one-factor model holding for data resulting from a dichotomization of underlying multivariate normal variables is a sufficient condition for the normal-ogive model with normally distributed θ. Equivalence was proved by Takane and de Leeuw (1987), who also addressed extensions to polytomous and pair-comparisons data. Earlier contributions to this domain were provided by Christoffersson (1975) and Muthén (1978).

Historically, factor analysis was developed to deal with problems of multidimensionality in test construction. Given the equivalence between the one-factor and normal-ogive models, it should not come as a surprise that work on problems of multidimensionality in IRT has also been stimulated by factor analysis. An important example is the work by McDonald (1967; this volume).

Introduction to Extensions of Dichotomous IRT

Although some IRT models for tests producing polytomous response data or for tests measuring more than one ability variable were published earlier (Bock, 1972; Rasch, 1961; Samejima, 1969, 1972), the majority of the models in this volume were produced in the 1980s and later. Several authors have provided illuminating reviews of these newer models. Masters and Wright (1984) reviewed extensions of the 1-PL or Rasch model, including the partial credit model, rating scale model, bionomial trials, and Poisson counts model. The main theme in their review was to show that some of these models can be obtained as generalizations or limiting cases of others.

Thissen and Steinberg (1984) reviewed a larger collection of models developed to deal with categorical data and provided a framework for classifying them. They also showed some ingenious reparametrizations through which some of the models could be obtained from others. The same theme was present in a review by Mellenbergh (1994), who showed how, via an adequate choice of a linking function and parameter structure, the majority of the extant models could be obtained from the generalized linear model (GLIM) (McCullagh and Nelder, 1989).

Mellenbergh (1995) classified the polytomous IRT models according to their response format. His main point of view is that both categorical and ordinal response variables can be split into dichotomies that can be combined in different ways to yield the various models. Important principles of combination for ordinal formats are Agresti's (1990, Sec. 9.3) principles of adjacent categories, cumulative probabilities, and continuation ratios. The same notion of underlying dichotomies was present in a review by van Engelenburg (1995), who considered these dichotomous as representations of the cognitive step an examinee needs to take to solve a test item.

The models in this volume are grouped into six different sections. The first section contains eight chapters on models for responses to items with a polytomous format or to open-ended items scored for the degree of completeness of the response. If the distractors of the item play a role in the solution process or the item formulates a problem consisting of different sequential components, the properties of these distractors or components must be parametrized in the model. All of the models in this section are unidimensional logistic models that have been extended to handle polytomous response data, either ordered or unordered. The pioneer in this area of model development is Fumiko Samejima with her graded response model, which could be applied to ordered categorical data. Variations on the graded response model of Samejima are the IRT models presented in the chapters by Andersen, Masters and Wright, Verhelst, Glas, and Vries, Tutz, and Muraki. For unordered categorical data, as might be reflected in the set of answer choices for a test item, the chapter by Bock describes his nominal response model. By scoring the answer choices to an item instead of simply scoring the item response as correct or incorrect, the potential was available for extracting more information about examinee ability. An extension of that model to handle the problem of guessing on multiple-choice items is described by Thissen and Steinberg.

The second section deals with models for items in which response time or numbers of successful attempts are recorded. Response time plays a role in tests with a time limit. Generally, the presence of a time limit requires response speed as a second ability in addition to response accuracy. Tests with a time limit can only be scored satisfactorily if the interaction of the two abilities on the probability of success on the item as well as the response time needed are carefully modeled. Both the chapters by Verhelst, Verstralen and Jansen and Roskam provide models for this interaction. If test items can be considered as parallel, it makes sense to record the number of successes in a fixed number of trials or the number of trials needed to complete a fixed number of successes. The chapter by Spray provides models for these cases that have been proven to be successful for analyzing data from psychomotor tests.

A persistent problem in the history of test theory has been the modeling of examinee performance when multiple abilities are required to complete items in a test. Traditionally, factor analysis was used to analyze the presence of multiple abilities and to reveal the structure of its impact on the responses to items in the test. IRT models for multiple abilities or for a single ability variable needed to solve multiple components or parts in the items are presented in the third section of this volume. This section of the volume contains the descriptions of several multidimensional IRT models. Reckase, McDonald, and Fischer provide parallel developments for the multidimensional logistic, normal-ogive, and linear logistic models, respectively. These are extensions of the corresponding unidimensional versions of the models (although the Reckase chapter extends the polytomous response version of the unidimensional logistic IRT model). Kelderman provides a loglinear version of a multidimensional IRT model for handling polytomous data. A very different type of multidimensional model is represented in the chapter by Zwinderman. Here, multidimensionality is accounted for in the data using a multiple regression model with observed scores used instead of latent variables as the predictors. Two additional models are introduced in Section 3 to provide for a psychological accounting of examinee item performance and items. The unidimensional linear logistic Rasch models of Fischer allow the difficulty parameter to be accounted for by cognitive components. With a better understanding of what makes items difficult, more systematic test construction becomes possible. Embretson's work with multicomponent response models permits a better understanding of examinee performance when item performance involves the completion of many small steps.

The models in Section 4 share the focus on response functions to describe the properties of the items with the other models in this volume. However, rather than defining these functions on a parameter structure to deal with the various factors that influence success on the items, it is the less demanding (nonparametric) shape of the functions that does this job. One of the first to pioneer nonparametric probabilistic models for dichotomous responses was Mokken. His chapter reviews the assumptions underlying his model and provides the statistical theory needed to work with it. The chapter by Molenaar follows the same approach as Mokken's but addresses the case of responses to items with a polytomous format. In the chapter by Ramsay, statistical techniques based on ordinal assumptions are presented to estimate response functions for dichotomous responses nonparametrically. Parametric models for item responses not only have to parametrize completely all factors with an impact on the response process but also must have a realistic functional form. An important distinction in this respect is the one between monotone and nonmonotone response functions. Items in an achievement test should have a monotonically increasing form. The more ability possessed by the examinee, the greater the probability of success. Items or statements in an instrument for the measurement of attitudes or values appear to miss this property. Models for such nonmonotone items are offered in the fifth section of this volume. The chapter by Andrich presents a hyperbolic cosine model to describe the probability of endorsement of statements in an instrument for the measurement of attitudes. The chapter by Hoijtink introduces a model based on the Cauchy density, which can be used as an alternative to Andrich's model.

The final section of this volume contains chapters with models accounting for various special properties of the response process. The multiple-group IRT model of Bock permits, for example, the group rather than the individual to become the unit of analysis. Such a shift may be important in program evaluation studies and educational accountability initiatives and several other applications. Better model parameter estimates is one of the advantages of this line of modeling. A unified treatment of IRT problems involving multiple groups is another. The chapter by Rost offers a description of a discrete mixture model that can be used to describe differences between subpopulations of examinees using different strategies to solve the items. Since these subpopulations are assumed to be latent, there is no need to identify the strategy used by the individual examinees to apply the model. Models to handle local dependencies in the data are provided by Jannarone. As these dependencies can be serious and lead to the rejection of the assumption of unidimensionality in the test data, it is likely these models will become more important in the future. To date, it has been common to simply disregard these dependencies. But it has been shown that failure to attend to these departures can lead to quite negative consequences (Fennessy, 1995). The models presented by Hutchinson allow for the fact that examinees may have partial information about the question asked in a test item. Each of these models is based on the same idea of an underlying continuous variable reflecting how much mismatch an examinee experiences between the question in the stem of the item and a specific alternative.

Future of Item Response Theory

As outlined in the introduction to this chapter, IRT started out as an attempt to adjust test scores for the effects of such nuisance properties of test items as their difficulty, discriminating power, or liability to guessing. Realistic models with parameters for each of these effects were needed. Owing to the success of these models, soon the need was felt to develop models for a larger collection of tests than standard tests with a single ability variable and a dichotomous response format. Such models had to deal, for example, with the presence of nuisance abilities, the role of time limits and speed, the possibility of learning during the test, the impact of properties of item distractors, or different cognitive strategies to solve test items. To date, many such models have been studied—this volume reviews 27 of them. The implicit goal in modern IRT research is to expand the class of models to cover response data from tests with any "natural" format. To realize this goal, deeper insights into response processes and the precise way item properties and human abilities interact are needed.

The important question can be raised whether, in so doing, psychometrics has not entered the domain of mathematical psychology. Indeed, it may seem as if IRT is about to give up its interest in measurement and adopt a new identity as an enthusiastic provider of psychological models for cognitive processes. Some of the families of models presented in this volume do not suggest any interest in estimating ability parameters at all but are presented mainly as vehicles for specifying competing hypotheses regarding, for instance, problem-solving strategies, the composition of skills, or cognitive growth. Pertinent examples are given in the chapters by Embretson, Fischer, Kelderman, and Fischer and Selinger in this volume. Rather than freeing ability measurements from the effects of nuisance variables, the interest of these authors has shifted to the behavior of the nuisance variables themselves. In our opinion, this development should not result in embarrassment but be welcomed as an unexpected bonus from the application of IRT models. Even when the attempt would be to do so, the emergence of this new branch of mathematical psychology would never be able to replace psychometrics as an independent discipline of measurement. The need for measurement models to solve the daily problems in, for instance, selection procedures, licensing, test equating, and monitoring achievement will remain, and improved models to deal with such problems are always welcome. If, in addition, such models also have a spin off to psychological theory, so much the better. They may help to better integrate measurement and substantive research—areas that to date have lived too apart from each other.

Owing to the attempts to make its model applicable under more "natural" conditions, IRT has met several of the technical problems inherent in statistical modeling that more experienced sciences have struggled with for a longer period of time. Large numbers of parameters and a more complicated structure may make a model realistic, but there always is a price to be paid.

First of all, if the number of parameters increases, less data per parameter are available and parameter estimates may show serious instability. The problem usually becomes manifest if estimates for the same parameters vary largely in value between applications. The problem is aggravated if the model has a tradeoff between some of the parameters in their effects on the response probabilities. As already discussed for the 3-PL model, this condition is reminiscent of the problem of multicollinearity in multivariate linear regression, and a prohibitively large amount of data may be needed to realize stable parameter estimates. For more complicated models than the 3-PL, serious versions of the problem can be expected to arise.

Another likely problem with complicated models is lack of identifiability of parameter values. Formally, the problem exists if different sets of parameter values in the model generate the same success probabilities for the examinees on the items. If so, estimation of model parameters cannot result in unique estimates. Practical experience with seemingly simple nonlinear models has taught us that such models often entail vexing identifiability problems. For several of the models in this volume, the issue of nonidentifiability has not been completely addressed yet, and practical criteria necessary and sufficient for identifiability are still lacking.

Even if unique parameters exist, attempts to estimate them may result in ML estimates which are not unique or cases in which no ML estimates exist at all. An example discussed earlier was the 3-PL model with known item parameters which for several data sets has a likelihood function with multiple local modes (Samejima, 1973; Yen et al., 1991). Necessary and sufficient conditions under which data sets produce unique parameter estimates are not yet known for the majority of the models in this volume. A favorable exception is the linear logistic model (LLTM) by Fischer [for the conditions, see Fischer (1983)].

Finally, if unique parameter estimates are known to exist, they may still be hard to calculate. As already mentioned, even a simple model like the Rasch or 1-PL model, numerical complexities had prevented the use of CML estimation for a long time until improved algorithms became available (Liou, 1994; Verhelst et al., 1984). For some of the models in this volume, numerical complexity can only be mastered using a modern technique such as the EM algorithm or one of the recent Monte Carlo methods for finding modes of posterior distributions or likelihood functions [for an introduction, see Tanner (1993)].

An important question, thus, is how far IRT modeling can go without loosing statistical tractability of its models. One obvious strategy is to work within families of distributions, which are known to have favorable properties and are flexible enough to fit empirical data. A well-explored example is the exponential family, which offers favorable minimal sufficient statistics for its parameters (Andersen, 1980). The advantage of these statistics is not so much the fact that they retain all information on the parameters in the sample but have a dimensionality much lower than the original data—a fact that may facilitate computation considerably. However, even within this family, application of advanced numerical procedures to estimate parameters may still be time consuming when applied to models for tests of realistic lengths (for an illustration, see Kelderman, this volume).

It is well known and often stated that models are always based on idealizations and will never completely fit reality. But even if this fact is taken into account, IRT may find that its goal of a realistic model for each possible test design cannot be fully realized because of technical restrictions. When this point will be reached is an issue that depends as much on the creativity of IRT researchers as on future developments in statistical theory and computing. It will be fascinating to follow IRT when models are further extended and refined in the decades to come.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York, NY: Wiley.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* **38**, 123–140.

Andersen, E.B. (1980). *Discrete Statistical Models with Social Science Applications.* Amsterdam, The Netherlands: North-Holland.

Baker, F.B. (1992). *Item Response Theory: Parameter Estimation Techniques.* New York, NY: Marcel Dekker.

Berkson, J.A. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association* **39**, 357–365.

Berkson, J.A. (1951). Why I prefer logits to probits. *Biometrics* **7**, 327–329.

Berkson, J.A. (1953). A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. *Journal of the American Statistical Association* **48**, 565–600.

Berkson, J.A. (1955). Maximum likelihood and minimum chi-square estimates of the logistic function. *Journal of the American Statistical Association* **50**, 120–162.

Binet, A. and Simon, Th.A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *l'Année Psychologie* **11**, 191–336.

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. and Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika* **46**, 443–459.

Bock, R.D. and Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika* **35**, 179–197.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika* **40**, 5–32.

Cox, D.R. (1958). *The Planning of Experiments.* New York, NY: Wiley.

de Leeuw, J. and Verhelst, N.D. (1986). Maximum-likelihood estimation in generalized Rasch models. *Journal of Educational Statistics* **11**, 183–196.

Engelen, R.H.J. (1989). *Parameter Estimation in the Logistic Item Response Model.* Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands.

Fennessy, L.M. (1995). *The Impact of Local Dependencies on Various IRT Outcomes.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Ferguson, G.A. (1942). Item selection by the constant process. *Psychometrika* **7**, 19–29.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern, Switzerland: Huber.

Fischer, G.H. (1983). Logistic latent trait models with linear constraints. *Psychometrika* **48**, 3–26.

Glas, C.A.W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika* **53**, 525–546.

Glas, C.A.W. (1989). *Contributions to Estimating and Testing Rasch Models.* Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands.

Haley, D.C. (1952). *Estimation of the Dosage Mortality Relationship When the Dose is Subject to Error* (Technical Report No. 15). Palo Alto, CA: Applied Mathematics and Statistics Laboratory, Stanford University.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (ed.), *Educational Measurement* (3rd ed., pp. 143–200). New York, NY: Macmillan.

Hambleton, R.K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer Academic Publishers.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Hattie, J. (1985). Assessing unidimensionality of tests and items. *Applied Psychological Measurement* **9**, 139–164.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika* **49**, 223–245.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh* **61**, 273–287.

Lazarsfeld, P.F. (1950). Chapters 10 and 11 in S.A. Stouffer et al. (eds.), *Studies in Social Psychology in World War II: Vol. 4. Measurement and Prediction.* Princeton, NJ: Princeton University Press.

Lazarsfeld, P.F. and Henry, N.W. (1968). *Latent Structure Analysis.* Boston, MA: Houghton Mifflin.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses.* New York, NY: Wiley.

Liou, M. (1994). More on the computation of higher-order derivatives of the elementary symmetric functions in the Rasch model. *Applied Psychological Measurement* **18**, 53–62.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs* **61** (Serial No. 285).

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

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Erlbaum.

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Masters, G.N. and Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika* **49**, 529–544.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* (2nd edition). London: Chapman and Hill.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monograph*, No. 15.

McDonald, R.P. (1989). Future directions for item response theory. *International Journal of Educational Research* **13**, 205–220.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin* **115**, 300–307.

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

Mislevy, R.L. (1986). Bayes modal estimation in item response theory. *Psychometrika* **51**, 177–195.

Molenaar, W. (1974). De logistische en de normale kromme [The logistic and the normal curve]. *Nederlands Tijdschrift voor de Psychologie* **29**, 415–420.

Molenaar, W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika* **55**, 75–106.

Mosier, C.I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review* **47**, 355–366.

Mosier, C.I. (1941). Psychophysics and mental test theory. II. The constant process. *Psychological Review* **48**, 235–249.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* **43**, 551–560.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology* **3**, 1–18.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4, pp. 321–333). Berkeley, CA: University of California.

Richardson, M.W. (1936). The relationship between difficulty and the differential validity of a test. *Psychometrika* **1**, 33–49.

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

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

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika* **38**, 221–233.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72–101.

Swaminathan, H. and Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics* **7**, 175–192.

Swaminathan, H. and Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika* **50**, 349–364.

Swaminathan, H. and Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika* **51**, 589–601.

Takane, Y. and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* **52**, 393–408.

Tanner, M.A. (1993). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* New York, NY: Springer-Verlag.

Thissen, D. (1982). Marginal maximum-likelihood estimation for the one-parameter logistic model. *Psychometrika* **47**, 175–186.

Thissen, D. and Steinberg, L. (1984). Taxonomy of item response models. *Psychometrika* **51**, 567–578.

Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* **16**, 433–451.

Thurstone, L.L. (1927a). The unit of measurements in educational scales. *Journal of Educational Psychology* **18**, 505–524.

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

Tsutakawa, R.K. (1984). Estimation of two-parameter logistic item response curves. *Journal of Educational Statistics* **9**, 263–276.

Tsutakawa, R.K. and Lin, H.Y. (1986). Bayesian estimation of item response curves. *Psychometrika* **51**, 251–267.

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

Urry, V.W. (1974). Approximations to item parameters of mental test models. *Journal of Educational Measurement* **34**, 253–269.

van Engelenburg, G. (1995). *Step Approach and Polytomous Items* (internal report). Amsterdam, The Netherlands: Department of Methodology, Faculty of Psychology, University of Amsterdam.

van der Linden, W.J. (1986). The changing conception of testing in education and psychology. *Applied Psychological Measurement* **10**, 325–352.

van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika* **47**, 123–139.

Verhelst, N.D., Glas, C.A.W. and van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly* **1**, 245–262.

Verhelst, N.D. and Molenaar, W. (1988). Logit-based parameter estimation in the Rasch model. *Statistica Neerlandica* **42**, 273–295.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurements* **5**, 245–262.

Yen, W.M., Burket, G.R. and Sykes, R.C. (1991). Nonunique solutions to the likelihood equation for the three-parameter logistic model. *Psychometrika* **56**, 39–54.