Measuring the Difference Between Two Models

Michael V. Levine, Fritz Drasgow, Bruce Williams,
Christopher McCusker, and Gary L. Thomasson
University of Illinois

Two psychometric models with very different parametric formulas and item response functions can make virtually the same predictions in all applications. By applying some basic results from the theory of hypothesis testing and from signal detection theory, the power of the most powerful test for distinguishing the models can be computed. Measuring model misspecification by computing the power of the most powerful test is proposed. If the power of the most powerful test is low, then the two models will make nearly the same prediction in every application. If the power is high, there will be applications in which the models will make different predictions. This measure, that is, the power of the most powerful test, places various types of model misspecification— item parameter estimation error, multidimensionality, local independence failure, learning and/or fatigue during testing—on a common scale. The theory supporting the method is presented and illustrated with a systematic study of misspecification due to item response function estimation error. In these studies, two joint maximum likelihood estimation methods (LOGIST 2B and LOGIST 5) and two marginal maximum likelihood estimation methods (BILOG and ForScore) were contrasted by measuring the difference between a simulation model and a model obtained by applying an estimation method to simulation data. Marginal estimation was found generally to be superior to joint estimation. The parametric marginal method (BILOG) was superior to the nonparametric method only for three-parameter logistic models. The nonparametric marginal method (ForScore) excelled for more general models. Of the two joint maximum likelihood methods studied, LOGIST 5 appeared to be more accurate than LOGIST 2B. *Index terms: BILOG; forced-choice experiment; ForScore; ideal observer method; item response theory, estimation, models; LOGIST; multilinear formula score theory.*

Statistical theory provides a most powerful statistical test for distinguishing between a pair of statistical models (Bickel & Doksum, 1977, section 6.1). The most powerful statistical test provides a common metric for comparing models. If the power of the most powerful statistical test is very low, then the models are very similar for most purposes. Under these circumstances, even if the most powerful statistical test is used, it is difficult to distinguish one model from the other. Therefore, typical applications will produce similar results, no matter which of the two models correctly describes the data. Thus, using the power of the most powerful statistical test to measure the difference between two psychometric statistical models is proposed.

The need for measuring the difference between models arises in many contexts, many of which concern model misspecification. For example, even if test data truly conform to a two-parameter logistic model, a one-parameter Rasch model may fit the data quite well if the item discriminations are generally close to 1 and trait levels are normally distributed (Lord, 1983). Similarly, if data are generated by a two-dimensional latent trait model with highly correlated traits, then a one-dimensional model also might describe the data well. If data fail to be locally independent because some examinees learn during testing (and the learning model is stated precisely enough to permit simulation), then the results presented below can be used to measure how well a particular latent trait model approximates the more complex model.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 3, September 1992, pp. 261-278
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/03261-18$2.15

The main advantage of having a common measure of the difference between models is that such a measure facilitates comparing misspecification in complex, poorly understood domains to misspecification in more familiar domains. For example, it has been observed (by the present authors and by Davey & Hirsch, 1991) that the model for a widely used two-dimensional test could be approximated with a one-dimensional model fairly well. Using the method presented below, the dimensionality misspecification can be related to test calibration error.

For the purposes of this paper, it is helpful to distinguish between statistical models and parametric models. A statistical model is specified when enough information is given to calculate ''manifest probabilities'' (Cressie & Holland, 1983; Lazarsfeld, 1959). In item response theory (IRT), statistical models generally are identified by specifying item response functions (IRFs) and a trait (θ) density. For any pattern, the manifest probability (the probability of sampling an examinee with that pattern) is simply the integral of the product of the pattern's likelihood function times the θ density.

A parametric model in IRT is a large set of statistical models, each of which has IRFs with a common parametric form. Thus, the Rasch model is a parametric model. A statistical model belonging to the Rasch model is obtained by specifying a vector of item difficulties and a θ distribution.

When the method proposed here is used to study misspecification due to parameter estimation error, it is often measuring the difference between two statistical models belonging to the same parametric model. The theoretical sections of this paper provide methods that allow comparisons of nonparametric models and non-latent-trait models that generally do not have a convenient parameterization. Those sections are primarily concerned with statistical models, which can be thought of as a list of manifest probabilities (see the discussion of ''standard multinomial forms,'' below).

The methods introduced are appropriate for comparing

1. Two models both belonging to the same parametric model, such as two three-parameter logistic parametric models with different item parameters and the same θ distribution;
2. IRT models belonging to different parametric models—for example, a Rasch model (i.e., with the θ distribution and all item difficulties specified) to a particular two-parameter logistic model; and
3. IRT models with process or other models that may not be considered IRT.

To compare two models, it is sufficient that both models be specified with enough detail to permit the calculation of the theoretical probabilities of every item response vector or, equivalently, to permit computer simulation. The method, in its present form, cannot be used to compare the Rasch parametric model with any other parametric model because the Rasch model cannot be simulated or used to calculate item response probabilities until a θ distribution and item difficulties are specified.

Unanticipated Applications, Model Differences, and Statistical Tests

When two statistical models predict approximately the same relative frequencies of item response patterns, it is difficult to determine whether the differences between the models should be ignored, especially when the intended applications of the models are diverse and not fully specified. For example, consider two logistic models differing only in their item parameters. Suppose Model A's parameters were estimated from a sample of examinees with very few minority examinees, and Model B's parameters were estimated from a sample containing a much larger proportion of minority examinees. A simulation study can be used to determine whether the proportion of minority examinees in a specified percentile range would be greater if Model A item parameter estimates or Model B item parameter estimates were used to compute a particular statistic that orders examinees.

Note that the outcome of the simulation study could be different if a different statistic was used to order examinees or if a different percentile range was selected. Thus, a simulation study can determine whether the difference between two models can be ignored for a particular problem, but it does not determine if model differences can be ignored in any other application.

This limitation on the interpretability of simulation studies is unfortunate because some testing programs use the same statistical model for many applications with a large number of examinees over a long period of time. Moreover, quite often at the time a model is selected, not all of the applications of the model are specified in sufficient detail to allow alternative models to be evaluated by simulation studies. Therefore, it is difficult or impossible to use such simulation studies to determine whether differences between two or more alternative models can be ignored.

Alternative statistical models can be compared in a way that is not specific to a single application. Classical statistical theory makes possible a ''worst case'' analysis that can be used to determine how discrepant the predictions of the models are in the application in which the differences between the models are the greatest. If small differences are obtained in this analysis, then small differences will be found in all other applications. The analysis also leads to a quantification of the differences between models and, in some situations, a measure of how well a model fits data.

The method depends on the following indirect reasoning: If there were an application in which Model A and Model B predict markedly different outcomes, this application could be used to construct a powerful statistical test of

H_A: sampled item response patterns are correctly described by Model A
against the hypothesis
H_B: sampled item response patterns are correctly described by Model B.

Therefore, if it can be shown that the most powerful statistical test has low power, there cannot be an application with markedly different outcomes. Using the methods described and illustrated in this paper, the power of the most powerful statistical test of H_A versus H_B is computed.

If the most powerful statistical decision-making procedure is no better than basing decisions on a random process, then Model A and Model B will not make very different predictions in any application. Conversely, if the most powerful statistical test is very powerful, then the test statistic can be used to specify a situation in which the models will predict clearly different outcomes. Between these two extremes, using the power of the most powerful statistical test as an index of the degree of difference between a pair of models is proposed.

Applying Signal Detection Theory to IRT

The similarity of a pair of psychological stimuli is sometimes measured by the relative ease with which a careful observer can distinguish between the stimuli in a reference experiment. Drawing on statistical decision theory, signal detection theory provides methods for quantifying the ease of discrimination. Green & Swets (1966) provided a clear introduction to signal detection theory that contains the first statement of a result used frequently in this paper.

In the two-alternative forced-choice experiment, an observer's ability to distinguish between a pair of stimuli is measured by how well the observer is able to guess the order in which the two stimuli have been presented. On approximately half of a large number of trials, Stimulus A precedes Stimulus B. On the remaining trials, Stimulus B precedes Stimulus A. The observer is informed that there will be no AA or BB trials, that the trials are independent, and that the probability of an AB trial equals the probability of a BA trial.

The observer is instructed to observe the stimuli as they are presented and then decide whether an AB pair or a BA pair was presented. Note that if the observer ignores the stimuli and randomly guesses, the observer will be correct 50% of the time. Thus, the smallest correct classification rate that should be obtained is .50. As the observer's ability to distinguish Stimulus A from Stimulus B increases, the proportion of correct answers will increase to a maximum value of 1.0.

To benchmark performance, the psychophysicist studies the behavior of an ''ideal observer''—that is, an observer making statistically optimal decisions. The ideal observer's success rate can be computed when detailed statistical models for Stimuli A and B are specified (details are given below). In other words, it is possible to compute a success rate that is actually achieved by a particular strategy and is at least as high as the success rate produced by every other decision strategy. Furthermore, signal detection theory provides a general efficient procedure (i.e., Green's theorem, which is given below) for approximating the optimal success rate.

Application to a Psychometric Problem

In this paper, the psychophysicist's paradigm and theoretical results are applied to the practical psychometric problem of quantifying the difference between two fully specified models. As noted above, the models may have the same or different parametric forms. The method requires only item response vector probabilities.

Consider an experiment in which a pair of simulated item response vectors is presented. One of the vectors, called the ''Model A response pattern'' or simply ''A,'' is obtained by simulating the first model. The ''B'' or ''Model B response pattern'' is obtained by simulating the second model. An unbiased coin is flipped to decide whether A or B is to be presented first to an observer. The hypothetical ''ideal observer'' attempts to use knowledge about the two models to decide whether the order is AB or BA. The hypothetical observer is ''ideal'' in the sense that the information about the models is used in a statistically optimal way: The observer uses a classification algorithm that has been proven to make correct decisions at least as often, asymptotically, as any competing algorithm.

One plausible measure of the dissimilarity of models is the success rate of an ideal observer (that is, an observer making statistically optimal decisions and thereby achieving the highest possible success rate). If the ideal observer's success rate is .50, then data cannot be used to distinguish between the models. In fact, the probability of every item response vector must be the same for both models. If there is even one pattern with higher probability according to one model, then the pattern could be used as an indicator of the model and a success rate greater than .50 would be achieved. Thus, success rates higher than .50 indicate models that can be differentiated; the higher the rate, the more reliably the models can be distinguished.

Now consider three models, for example, Model A, Model B, and Model C, that belong to a common parametric family and differ only in their item parameters. If the percent correct for Model A versus Model B is close to 1.0, and the percent correct for Model A versus Model C is close to the chance level .50, then it is much easier to distinguish B from A than to distinguish C from A. Thus, Model C approximates Model A better than does Model B. Therefore, using the probability of a correct discrimination in the two-alternative forced-choice experiment is proposed to indicate the degree of similarity of a pair of item response models.

Use of Percent Correct Discrimination as a Measure of Model Similarity

Estimated versus true item parameters. In many practical situations, an item response model is used that is known a priori to be somewhat false. Consider a situation in which item parameters have been estimated with error, and assume that the estimation error is the only source of model misspecification. The ideal observer method can be used to determine whether the estimated parameters are indistinguishable—for all practical purposes—from the true parameters, given that the item response model is correctly specified and a calibration sample of specified size is drawn. When one application is specified, a simulation study generally can be used to analyze the consequences of estimation error or some other form of misspecification. The ideal observer method permits many applications to be considered simultaneously.

The ideal observer method could be used to evaluate estimated item parameters in the context of a simulation study such as the following. Begin by selecting a representative set of item parameters that serve as the true item parameters, which will be denoted Model A. These parameters are used to generate a sample of the size hypothesized to be adequate for item calibration. Parameters are estimated from the simulated calibration sample. Model B contains the estimated item parameters. Next, both the simulated and estimated item parameters are used to generate a sample of response patterns. Using the methods described below, the success rate of the ideal observer in distinguishing Model A response patterns (i.e., patterns generated from the true item parameters) and Model B response patterns (i.e., patterns generated from the estimated item parameters) in the two-alternative forced-choice experiment is calculated. If the ideal observer's success rate is at the chance level (.50), then Models A and B make identical predictions about the frequencies of each response pattern and, consequently, true and estimated item parameters are equivalent. To the extent that the results from the simulation study can be generalized to real data (e.g., if a model's assumptions about dimensionality are correct for a dataset), test calibrations with samples of the size used in the simulation are justified.

Comparison of two best-fitting parametric models. Many parametric IRT models have been proposed. Suppose two of these models have been fitted to a dataset and it is important to determine the extent to which the fitted models can be distinguished. The ideal observer method allows the relation between the two models to be measured, but does not address which model is closer to the "true" underlying model.

For example, consider two unidimensional parametric models for polychotomously scored item responses. Let Model A denote Samejima's (1979) multiple-choice model and let Model B denote Thissen & Steinberg's (1984) multiple-choice model. Suppose both models were fitted to a mathematics test using a large sample so that estimation errors can be ignored (this sample size may have been determined by the process described above). The degree to which Model A differs from Model B can be determined by a simulation study similar to that described above based on a large sample of both Model A and Model B response patterns. The ideal observer's success rate in distinguishing Model A response patterns from Model B response patterns in the two-alternative forced-choice experiment again is used to determine the degree of difference between the models. If the ideal observer's success rate is very close to .50, then the models are virtually identical; if the success rate is close to 1.00, the two models provide very different representations of the mathematics test. This approach to model comparison also can be applied to unidimensional approximations of multidimensional models, to logistic approximations of non-logistic models, and to the use of standard IRT models for approximating cognitive process models.

Scrambled forms of a test. Suppose Test B contains the same items as Test A, but the items are arranged in different orders or are presented through a different medium. Suppose further that data have been collected for both tests, item parameters have been estimated, and some item parameter estimates have been found to differ beyond the degree expected on the basis of chance (perhaps by the method outlined by Lord, 1980, chap. 14). The effects of the set of item parameter estimates on θ estimation (e.g., for the expected a posteriori estimates of Bock & Mislevy, 1982), appropriateness measurement (Levine & Rubin, 1979), or other analyses of an individual examinee's responses is a question of the comparison between models.

Typically, comparisons of this kind are approached by simulating each of the two models, selecting an application (e.g., θ estimation), selecting a θ measure [e.g., maximum likelihood estimate

(MLE) of θ], and determining by monte carlo methods whether examinees are likely to be found who have substantially different MLE $\hat{\theta}$s when the two sets of item parameters are used. A problem with this approach is that the outcome can depend on the statistic selected.

The ideal observer method provides a powerful means for comparing the two sets of parameter estimates. Item parameter estimates from Test A can be used as Model A, estimates from Test B can be used as Model B, and the simulation study previously described can be applied. If the correct classification rate of the ideal observer is near .50, then either set of item parameter estimates can be used when analyzing the responses of individual examinees. This conclusion is appropriate because the classification procedure of the ideal observer is optimal; MLE $\hat{\theta}$ or any other trait estimate cannot show larger differences. Other applications, such as appropriateness measurement, also will be unaffected by the choice of item parameter estimates. The results will then generalize across applications.

Computer program enhancements. Computer programs used to estimate the parameters of an IRT model incorporate a complex mixture of statistical theory, numerical analysis, and heuristics. Programs are often revised to improve performance. The ideal observer method provides an effective means for comparing two versions of an estimation algorithm.

In this situation, a representative set of item parameters can serve as Model A, parameter estimates obtained with the old algorithm can serve as Model B, and parameter estimates obtained with the new version of the program can serve as Model C. The new version of the program is an improvement if the ideal observer's correct classification rate is lower when comparing Models A and C than when comparing Models A and B.

Comparing Models With the Ideal Observer Method

To compare two item response models, consider their standard multinomial forms. The multinomial form of an item response model is a table listing all the item response patterns and their probabilities. Two item response models are considered equivalent if they have the same multinomial form.

Model equivalence is important because equivalent models have identical distribution functions for any statistic computed from the item responses. This is easily proven by expressing the distribution function

$$P\{X \leq x\} \tag{1}$$

of a statistic X in terms of sums of pattern probabilities. Thus, for each number x

$$P\{X \leq x\} = P(\mathbf{u}_1^*) + P(\mathbf{u}_2^*) + , \ldots , + P(\mathbf{u}_M^*) \quad , \tag{2}$$

where $\mathbf{u}_1^*, \mathbf{u}_2^*, \ldots, \mathbf{u}_M^*$ is an enumeration of the item response patterns having statistic values $X(\mathbf{u}^*)$ less than or equal to x. Because equivalent models have identical probabilities of response patterns, the distribution functions of any statistic X also must be identical.

Comparing multinomial forms for different models is a nontrivial, practical, and theoretical problem, which is discussed below. The ideal observer method is one approach to this problem. Two response patterns, $\mathbf{u}_1^*$ and $\mathbf{u}_2^*$, are presented to the observer. To specify an optimal strategy, the probabilities $P_A(\mathbf{u}_1^*)$ and $P_B(\mathbf{u}_1^*)$ of response pattern $\mathbf{u}_1^*$ are used to compute the likelihood ratio statistic ℓ by

$$\ell(\mathbf{u}_1) = P_A(\mathbf{u}_1^*)/P_B(\mathbf{u}_1^*) \quad . \tag{3}$$

The likelihood ratio $\ell(\mathbf{u}_2^*)$ is computed in the same way for pattern $\mathbf{u}_2^*$. The following decision rule maximizes the probability of correctly classifying response patterns:

Respond AB if $\ell(\mathbf{u}_1^*) \geq \ell(\mathbf{u}_2^*)$; otherwise respond BA. (4)

The Neyman-Pearson lemma (Kendall & Stuart, 1979; Lehmann, 1959) can be used to show that this strategy maximizes the percent of correct classifications.

The Yes-No Experiment

An experimental procedure from the signal detection literature that complements the two-alternative forced-choice experiment and facilitates the calculation of the ideal observer's correct classification rate is the yes-no experiment (Green & Swets, 1966, chap. 2). In this experiment, the observer is presented with a single stimulus and must decide whether it was generated using Model A or Model B. According to the Neyman-Pearson lemma, an optimal statistical test is provided by the likelihood ratio statistic. The likelihood ratio decision rule with criterion k is:

Choose Model A if $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*) > k$; otherwise choose Model B. (5)

The Neyman-Pearson lemma guarantees that the test with criterion k maximizes power, defined as

$$P\{\text{Model A is chosen} \mid \text{Model A is correct}\}, \qquad (6)$$

among all tests with the same level of significance, as determined by

$$P\{\text{Model A is chosen} \mid \text{Model B is correct}\} \qquad (7)$$

as the likelihood ratio test with criterion k.

A summary of the performance of the ideal observer in the yes-no experiment is provided by a table listing the probabilities of correct decisions and the associated error rate for the various values of k. A plot of the values in this table is called a receiver operating characteristic (ROC) curve of the likelihood ratio statistic. The ROC curve graphically displays the maximum correct classification rate at each level of significance. Specifically, the ROC curve is the set of points (x_k, y_k) in the unit square where

$$x_k = P\{P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*) > k \mid \text{Model B}\} \qquad (8)$$

and

$$y_k = P\{P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*) > k \mid \text{Model A}\}. \qquad (9)$$

There is one point for each k. Nearly equivalent models will have a ROC curve with points slightly above the 45° line $f(x) = x$.

Green (see Green & Swets, 1966) discovered an important relation between the accuracy of the ideal observer in the two-alternative forced-choice experiment and in the yes-no experiment. Green's theorem states that the area under the curve formed by connecting consecutive points on the ideal observer's ROC curve in the yes-no experiment is equal to the ideal observer's correct classification rate in the two-alternative forced-choice experiment. Thus, the ideal observer's performance in the two-alternative forced-choice experiment can be predicted from the yes-no experiment.

To quantify the difference between models, the present authors prefer the two-alternative forced-choice experiment to the yes-no experiment, because the symmetry of the two-alternative forced-choice experiment reduces the classification rate to a single number. In contrast, the entire ROC curve is needed to describe the ideal observer's performance under various conditions in the yes-no experiment.

The yes-no experiment is nonetheless useful for several purposes. First, it can be used to predict the performance of the ideal observer in the asymmetric situation in which response patterns from one of the models are presented less frequently than response patterns from the other model. The yes-no ROC curve also indicates conditions under which the ideal observer method fails (this is further discussed below). Finally, concerns about sampling variability in monte carlo studies have led to an experimental procedure in which the yes-no paradigm is used, a ROC curve is constructed for the likelihood ratio, and Green's theorem is used to determine the ideal observer's correct classification rate in the two-alternative forced-choice experiment.

Monte carlo methods can be used to estimate the yes-no ROC curve. Candell (1988) studied how well empirical ROC curves approximate theoretical ROC curves. He obtained very accurate recovery with samples of 2,000 Model A response patterns and 4,000 Model B response patterns. Small bias in the monte carlo estimates of points on the ROC curve should result in small bias in the estimate of the two-alternative forced-choice optimal classification rate, but this has not been investigated systematically.

Summary of Computational Procedures

Let Model A and Model B denote the two models under consideration. Large samples of response patterns are simulated for Model A (e.g., $N_{\mathrm{A}} = 3{,}000$) and for Model B. Let $P_{\mathrm{A}}(\mathbf{u}^*)$ and $P_{\mathrm{B}}(\mathbf{u}^*)$ denote the probability of pattern $\mathbf{u}^*$ given Model A and Model B, respectively.

Compute the likelihood ratio

$$\ell = P_{\mathrm{A}}(\mathbf{u}^*) / P_{\mathrm{B}}(\mathbf{u}^*) \tag{10}$$

for each Model A and Model B response pattern. The set of likelihood ratios for Models A and B then are used to determine a ROC curve (see Hulin, Drasgow, & Parsons, 1983, pp. 131–135 for an example and detailed description of the construction of a ROC curve). Finally, the area under the ROC curve is used as the estimate of the ideal observer's correct classification rate.

Empirical Studies

Overview

Four simulation studies were performed to illustrate the ideal observer method. The simulation studies shared a number of features, including: (1) a unidimensional latent trait space and locally independent item responses; (2) dichotomously scored item responses; and (3) 30 items and 3,000 response patterns per simulated test. The studies varied in the distribution from which θs were sampled and the shapes of the IRFs.

The studies used the same experimental procedure, which is depicted in the flow chart shown in Figure 1. Each study began by specifying a θ distribution and selecting a set of IRFs, as shown in Table 1. The four studies were designed to simulate situations in which all standard parametric assumptions are satisfied, situations with modest violations of standard assumptions, and situations with moderately severe violations.

In the next step of each study, a test calibration sample of $N = 3{,}000$ simulated response patterns was generated using the simulation parameters. Item parameters then were estimated by four methods, using default convergence criteria. The methods included:

1. Estimation of the item parameters of the three-parameter logistic model by marginal maximum likelihood. PC-BILOG (Mislevy & Bock, 1984a) was used to obtain marginal MLEs. In preliminary analyses, biased estimates of item difficulty and item discrimination parameters were observed with the default value of 10 quadrature points. Increasing the number of quadrature points to 30 eliminated the bias; therefore, the larger number of quadrature points was used in subsequent

Figure 1
Flow Chart for the Simulation Studies

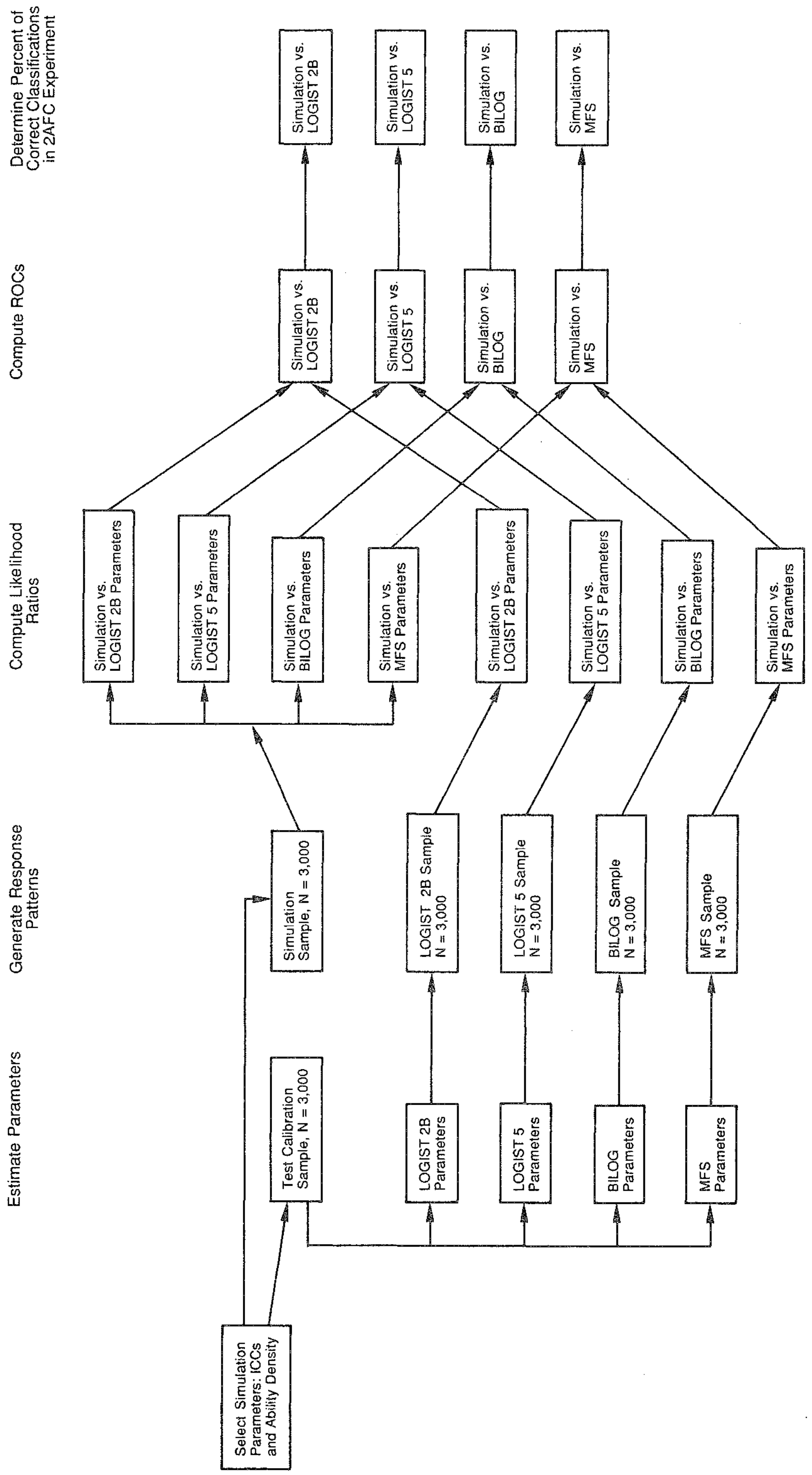

Estimate Parameters
Generate Response Patterns
Compute Likelihood Ratios
Compute ROCs
Determine Percent of Correct Classifications in 2AFC Experiment
Select Simulation Parameters: ICCs and Ability Density
Test Calibration Sample, N = 3,000
LOGIST 2B Parameters
LOGIST 5 Parameters
BILOG Parameters
MFS Parameters
Simulation Sample, N = 3,000
LOGIST 2B Sample N = 3,000
LOGIST 5 Sample N = 3,000
BILOG Sample N = 3,000
MFS Sample N = 3,000
Simulation vs. LOGIST 2B Parameters
Simulation vs. LOGIST 5 Parameters
Simulation vs. BILOG Parameters
Simulation vs. MFS Parameters
Simulation vs. LOGIST 2B Parameters
Simulation vs. LOGIST 5 Parameters
Simulation vs. BILOG Parameters
Simulation vs. MFS Parameters
Simulation vs. LOGIST 2B
Simulation vs. LOGIST 5
Simulation vs. BILOG
Simulation vs. MFS
Simulation vs. LOGIST 2B
Simulation vs. LOGIST 5
Simulation vs. BILOG
Simulation vs. MFS

Table 1
θ Distributions and IRFs Used in the Simulation Studies

Study	θ Distribution	IRF
1	N(0,1)	3PL
2	Mixture	3PL
3	Mixture	POLY8, Intact Test
4	Mixture	POLY8, Selected Items

analyses.

2. Item and person parameters of the three-parameter logistic model were estimated by joint maximum likelihood estimation using both LOGIST Version 2B (Wood, Wingersky, & Lord, 1976) and LOGIST Version 5 (Wingersky, Barton, & Lord, 1982) with default parameters.
3. IRFs were estimated by nonparametric marginal maximum likelihood estimation using the program ForScore (Williams & Levine, in preparation), which uses maximum likelihood formula scoring (MFS, also called multilinear formula scoring) and nonparametric IRFs (details of this analysis are in the Appendix).

Generation of Response Patterns

After obtaining item parameter estimates, a second sample of $N = 3{,}000$ was generated from the simulation parameters (a second sample was created to avoid artifacts due to overfitting). Samples of $N = 3{,}000$ also were generated with each of the four sets of estimated item parameters. The simulation θ distribution was used with item parameter estimates from the parametric estimation methods to generate response patterns. In contrast, the θ density estimated in the first phase of the MFS analysis was used in conjunction with the nonparametric IRFs. Nonparametric IRFs have meaning only in relation to some particular θ density; because the MFS IRFs were estimated in reference to an estimated density, it was necessary to use this estimated density in all subsequent analyses. This situation is the nonparametric analogue of the parametric model scaling indeterminacy that is often resolved by standardizing θs.

Likelihood Ratios

In the next stage of each simulation study, likelihood ratios were computed. Four likelihood ratios $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*)$ were computed for each response pattern from the second sample generated with the simulation parameters. For each of these likelihood ratios, $P_A(\mathbf{u}^*)$ refers to the probability of $\mathbf{u}^*$ computed with simulation parameters. This is the "true" probability of these $\mathbf{u}^*$ because they were generated with the simulation parameters. Four additional probabilities $P_B(\mathbf{u}^*)$ also were computed for the response patterns generated from the simulation parameters. Specifically, $P_B(\mathbf{u}^*)$ was computed using item parameter estimates obtained from each of the estimation methods described previously (i.e., BILOG, LOGIST 2B, LOGIST 5, ForScore). Then four likelihood ratios were formed: (1) $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*)$ computed with LOGIST 2B item parameter estimates; (2) $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*)$ computed with LOGIST 5 item parameter estimates; (3) $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*)$ computed with BILOG item parameter estimates; and (4) $P_A(\mathbf{u}^*)/P_B(\mathbf{u}^*)$ computed with ForScore IRF estimates.

A single likelihood ratio was computed for each response pattern generated by each of the sets of estimated item parameters. The probability $P_A(\mathbf{u}^*)$ in the numerator was computed using the simulation parameters. For the response patterns generated with LOGIST 2B item parameter estimates, the probability $P_B(\mathbf{u}^*)$ in the denominator was evaluated using the LOGIST 2B item parameter estimates. In this case, $P_B(\mathbf{u}^*)$ is the "true" probability because the model used to generate a response pattern (the LOGIST 2B item parameter estimates) was identical to the model used to compute its probability. Likelihood ratios were computed analogously for the response patterns generated from the LOGIST 5, BILOG, and ForScore estimated IRFs.

The marginal probabilities $P_A(\mathbf{u}^*)$ and $P_B(\mathbf{u}^*)$ required for the likelihood ratio were computed by

$$\int \prod \{[P_i(t)]^{u_i}[1 - P_i(t)]^{1-u_i}\} f(t)\, dt \quad , \tag{11}$$

where $P_i(t)$ was evaluated using simulation parameters or estimated parameters, respectively. The simulation density (shown in Table 1) was used for f when computing $P_A(\mathbf{u}^*)$. The simulation density used to generate the response patterns also was used in the calculation of $P_B(\mathbf{u}^*)$ for BILOG and LOGIST item parameter estimates. The density estimated by MFS was substituted for f when $P_B(\mathbf{u}^*)$ was evaluated with MFS IRFs. The integral in Equation 11 was approximated by numerical integration with 61 quadrature points evenly spaced between -3 and $+3$.

ROC Curves and Classification Rates

The likelihood ratios for response patterns generated from the simulation parameters were ordered from largest to smallest; likelihoods for the response patterns generated from each of the four types of item parameter estimates were similarly ordered. Hit rates, false positive rates, and ROC curves were computed for each of the four types of parameter estimates, and the areas under the four ROC curves were determined by numerical methods. Three replications were obtained in each of the four studies.

Study 1

This study was concerned with whether a sample of $N = 3{,}000$ is adequate for estimating the item parameters of a 30-item test when standard parametric assumptions are satisfied. It also examined how much is sacrificed by using a nonparametric model under conditions that are ideal for calibration with the three-parameter logistic model.

Method. The θ distribution in Study 1 was the standard normal. Estimates of the parameters of the three-parameter logistic model obtained by Mislevy & Bock (1984b) for the Arithmetic Reasoning subtest of the Armed Services Vocational Aptitude Battery were used as the simulation item parameters.

Results. The correct classification rates of the four estimation methods for each of the three replications are provided in Table 2, which shows that BILOG was very effective when all assumptions were satisfied. The most powerful test for differentiating response patterns generated by the simulation item parameters and response patterns generated by estimated item parameters could correctly classify on average less than 55% of such pairs. ForScore was surprisingly successful. The mean correct classification rate for this method was less than 2% higher than the rate of BILOG. The mean correct classification rate of ForScore was 2% less than joint estimation with LOGIST 2B. The classification rates for LOGIST 2B and LOGIST 5 were fairly similar. These rates were higher than the classification rates for the other two methods.

Discussion. Prior to Study 1, no hypotheses were made about the magnitudes of correct classification rates for the different estimation methods. Marginal estimation for the correct parametric model was expected to provide the lowest correct classification rate, but no strong conviction was held that this rate would be approximately 55% instead of, say, 85%. Study 1 provided strong support for the use of marginal maximum likelihood and the BILOG computer program when the assumptions are met. Study 1 also demonstrated that marginal estimation of nonparametric MFS IRFs, as

Table 2
Estimated Correct Classification Rates for the Ideal Observer
When Parameters Were Estimated by Four Methods

Study and	Estimation Method			
Replication	LOGIST 2B	LOGIST 5	BILOG	ForScore
Study 1				
1	.587	.581	.556	.571
2	.588	.579	.543	.566
3	.587	.586	.548	.564
Mean	.587	.582	.549	.567
Study 2				
1	.586	.582	.558	.563
2	.576	.578	.545	.556
3	.593	.578	.544	.569
Mean	.585	.579	.549	.563
Study 3				
1	.620	.609	.572	.547
2	.601	.606	.571	.565
3	.589	.601	.577	.566
Mean	.603	.605	.573	.559
Study 4				
1	.648	.612	.582	.580
2	.643	.601	.588	.564
3	.641	.626	.586	.567
Mean	.644	.613	.585	.570

implemented by ForScore, was also effective. Despite vastly increasing the set of permissible IRFs by allowing more general shapes than the three-parameter logistic model, estimation accuracy was only slightly less than the best parametric methods.

Study 2

Study 2 was concerned with whether realistic violations of the θ distribution assumption have deleterious effects on marginal maximum likelihood estimation for a parametric model. Neither LOGIST 2B nor LOGIST 5 makes any assumption about the θ distribution and, consequently, neither should suffer when the θ density is not standard normal. Similarly, no decrement in the performance of ForScore is expected to result from changes in the θ distribution, because the θ density is estimated in the first step of the MFS analysis.

Method. A situation was simulated in which the overall population consisted of two subpopulations with between-group differences—a situation commonly observed with standardized tests. Therefore, a mixture of two normal distributions was used: θ was sampled from a $N(.2,.917^2)$ distribution with probability .8 and from a $N(-.8,.917^2)$ distribution with probability .2. The mean and variance of the marginal distribution formed by this mixture were 0 and 1, respectively. The IRFs used in Study 2 were identical to those used in Study 1.

Results. The results from Study 2 were similar to those from Study 1 (see Table 2). As expected, estimation for the nonparametric method and the two versions of LOGIST was unaffected. Table 2 clearly shows that BILOG was similarly unaffected, even though a mixture of normal distributions was used to sample θs rather than a standard normal.

Discussion. In specifying the mixture distribution, a situation was considered in which the overall population consisted of a majority group and a minority group that had a difference in means of approximately one standard deviation and had normal within-group distributions. Despite sampling θs from two relatively divergent distributions, the mixture distribution was surprisingly similar to the standard normal distribution. Consequently, BILOG's performance was virtually unaffected by violation of its assumption about the θ distribution.

A distribution more radically different from the standard normal could have been used in Study 2. The mixture of normals was selected after considerable deliberation, because it seemed to be the most appropriate simulation of psychological aptitude and achievement tests. Of course, substantially non-normal distributions may arise with other latent variables. Investigations of BILOG's performance under these conditions is a matter for future research.

Study 3

This study concerned the effects of mild violations of the assumed form of IRFs on estimation accuracy of parametric IRT models. Because the nonparametric model does not assume any specific mathematical form for IRFs, its estimation algorithm should be unaffected when curves other than three-parameter logistic ogives are used to determine response probabilities.

Method. Sympson (1988) provided item parameter estimates obtained by his POLY program from an experimental 35-item test of word knowledge. The first 5 items were omitted in order to hold test length constant at 30 items across all studies reported here. POLY performs polychotomous test analysis. However, for the studies presented here, incorrect response categories were aggregated to form one incorrect category in order to maintain dichotomous scoring across all studies.

Sympson's (1988) model incorporates a very general function for modeling the probability of correct responses. To examine departures from the three-parameter logistic ogive, the weighted (by the normal density) least squares three-parameter logistic approximation to each of the 30 polyweight IRFs (after rescaling Sympson's percentile rank metric to the θ metric of IRT) was determined. Visual comparisons of Sympson's IRFs and corresponding three-parameter logistic approximations revealed 19 items that were almost perfectly fitted by three-parameter logistic curves, 5 items with slight departures, and 6 items with moderate differences between pairs of curves. Figure 2a shows an item classified as being slightly different, and Figure 2b shows an item classified as being moderately different. The θ distribution used in Study 3 was the same mixture of normal distributions used in Study 2.

Results. Table 2 shows that LOGIST and BILOG had modest decrements in performance due to the misspecification of IRFs. ForScore showed very little difference in Study 3 from its performance in Studies 1 and 2.

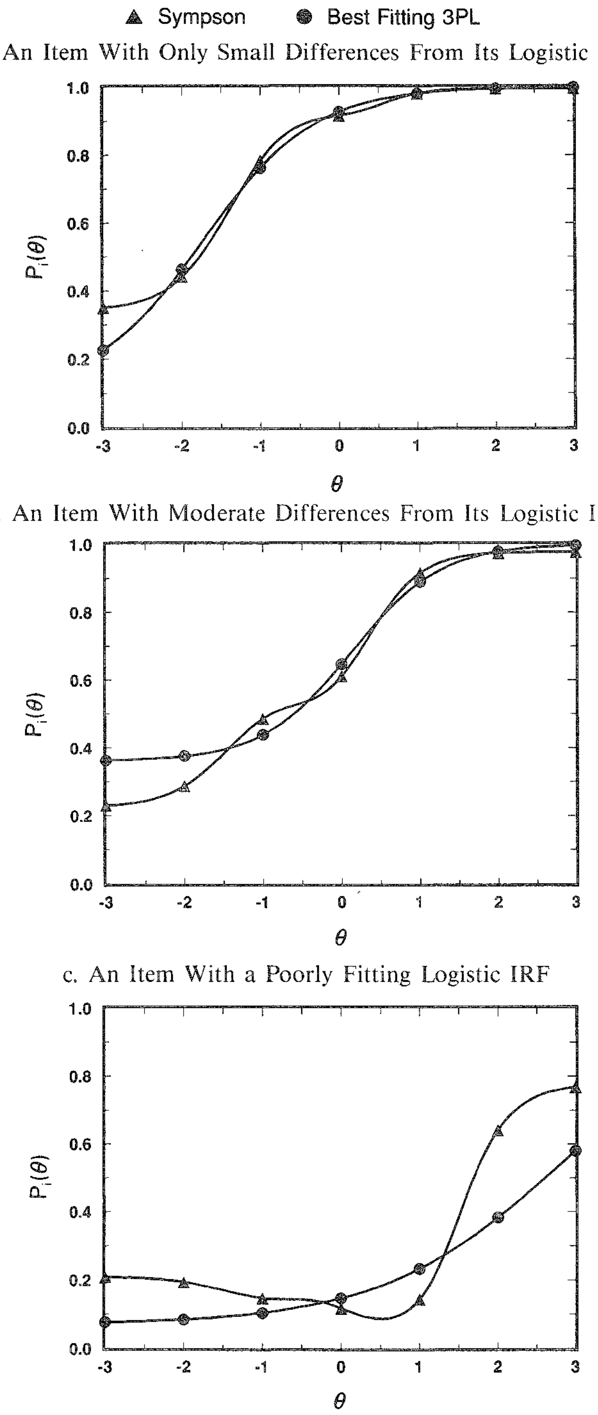
Discussion. Note that the mild violations of parametric assumptions introduced in Study 3 reversed the order of the correct classification rates for marginal maximum likelihood for the three-parameter logistic model (BILOG) and marginal maximum likelihood for the nonparametric model (ForScore). Both approaches seem quite satisfactory for practical use with such data.

Study 4

This study examined the effect of moderate violations of the assumption that IRFs are three-parameter logistic ogives.

Method. Sympson (1986) also provided polyweight item parameters for 86 additional items. The entire set of 121 items was inspected. The polyweight IRFs were compared to the best fitting three-parameter logistic curves (obtained by the same weighted least squares method used in Study 3), and the 30 items that appeared most poorly fit by three-parameter logistic ogives were selected. The item fit for many of these items was similar to the items shown in Figures 2a and 2b. An item with one of the worst fitting three-parameter logistic curves is shown in Figure 2c. The θ density used in Study

Figure 2
Polyweight Items and Their Approximating Three-Parameter Logistic IRF



a. An Item With Only Small Differences From Its Logistic IRF

b. An Item With Moderate Differences From Its Logistic IRF

c. An Item With a Poorly Fitting Logistic IRF

4 was the mixture of normal distributions used in Studies 2 and 3.

Results. Table 2 shows that the moderate model misspecification created for Study 4 had a rather large impact on LOGIST 2B. In contrast, LOGIST 5 and BILOG were more robust. ForScore was the most effective estimation method in the nonparametric setting investigated in Study 4.

Discussion. Study 4 confirmed the hypothesis that estimation accuracy of the nonparametric method would be unaffected by IRFs with shapes that differed from the three-parameter logistic ogive.

Discussion and Conclusions

There are several reasons why the ideal observer method is preferred to some of the simpler alternatives. First, a straightforward inspection of the corresponding pairs of pattern probabilities for two models is impractical because there are too many pattern probabilities to consider. For example, with a 40-item test there are approximately 1.099×10^{12} pattern probabilities. Thus, except for very short tests, direct comparison is not possible.

The ideal observer method computes a complicated function of a pair of multinomial forms. Because a multinomial form is simply a vector of pattern probabilities, using one of the standard metrics for measuring the distance between vectors—such as the Euclidean distance (i.e., the square root of the sum of squared differences between pattern probabilities) or the maximum of the absolute values of the differences between corresponding pairs of pattern probabilities—has been suggested. However, the standard metrics fail to adequately measure the difference between clearly distinct models, because vectors of pattern probabilities can be very close geometrically but very different in terms of the relative frequencies of response patterns. For example, for a 40-item test, half of the 10^{12} patterns must have probabilities smaller than a billionth. Consequently, pattern probabilities can agree to 10 decimal places for easily distinguished models. For additional details, see Levine, Drasgow, Williams, McCusker, & Thomasson (1992).

For very short tests and very large test administrations, the ideal observer method can measure goodness of fit. The ideal observer method can be used to measure the "relative" goodness of fit of parametric and nonparametric models that are unidimensional or multidimensional. The short test length and large administration also permit consideration of the "absolute" goodness of fit because a large sample of the administration can be used to estimate accurately the probabilities in the multinomial form and then define a psychometric model. To measure absolute goodness of fit, the ideal observer can be given the task of deciding which of a pair of item response vectors was taken from a hold-out sample and which was generated using one of the fitted models.

It is important to note two limitations of the ideal observer method. First, it is difficult to compare classification rates across simulations of tests of different lengths. Holding model misspecification constant, longer tests provide more information to the likelihood ratio in Equation 10, and therefore allow more accurate classification. Second, the ideal observer method treats each response pattern equally; this can lead to unsatisfactory results when some response patterns are highly diagnostic and their misclassification can lead to very adverse consequences. Levine et al. (1992) provide a detailed example of this latter problem and an additional discussion of the method.

In these studies, small but consistent advantages were observed for marginal maximum likelihood estimation. This finding corroborates results obtained by Drasgow (1989), who compared marginal MLEs of the item parameters of the two-parameter logistic model with joint MLEs. Less bias, smaller observed standard errors, and more accurate estimates of standard errors were found for the marginal estimates.

Table 2 also shows small but consistent advantages for LOGIST 5 relative to LOGIST 2B, but the two versions of LOGIST do not scale their parameter estimates in reference to a standard normal θ distribution. The simulation studies nonetheless used the standard normal density in Equation 2 and, therefore, the two versions of LOGIST were penalized to some degree. To address this concern, the θ density for each of the 12 test calibration samples was estimated using the LOGIST 5 item parameter estimates. The density estimates were based on Levine's (1989) MFS theory; ForScore (Williams & Levine, in preparation) was used for the estimation. This analysis explicitly placed the estimated densities in the scales established by the LOGIST 5 item parameter estimates.

After estimating θ densities, classification rates were recomputed for LOGIST 5. In each case, the θ density estimated in the scale of the parameter estimates was used in Equation 2 to compute the LOGIST 5 marginal probabilities. The classification rates were: .566, .561, and .559 for Study 1 (mean = .562); .578, .555, and .569 for Study 2 (mean = .567); .580, .585, and .578 for Study 3 (mean = .581); and .589, .585, and .606 for Study 4 (mean = .593). Thus, the differences between the standard normal density and the scale implicitly defined by LOGIST 5 were large enough to improve the correct classification rates by approximately .02. Using the estimated density reduced the differences between marginal and joint estimation (by approximately 50%), but did not change the overall conclusion that marginal estimation is preferable to joint estimation.

A second important result shown in Table 2 is the surprising accuracy of estimation for Levine's (1984, 1989) MFS theory. Even in the situation most favorable for marginal maximum likelihood estimation of the parameters of the three-parameter logistic model, ForScore (Williams & Levine, in preparation) yielded estimated IRFs that were only slightly less accurate than the curves obtained by BILOG. This result is rather surprising because the set of IRFs allowed by the MFS model is vastly larger than the set of three-parameter logistic IRFs.

In Study 2, BILOG was robust to violations of the assumption that θ followed a standard normal distribution. It was found to be less robust to misspecification of the IRFs in Studies 3 and 4. In contrast, the estimation accuracy of the nonparametric model was found to be relatively unaffected by IRFs with shapes that differed from three-parameter logistic ogives. This latter finding shows that ForScore has realized an important objective of the nonparametric item response theories.

Finally, the results of Studies 1 through 4 suggest the following strategy for IRT analyses of tests and scales. In the early stages of item development and pretesting, IRFs should be estimated by a nonparametric method such as ForScore. This analysis will identify miskeyed or otherwise flawed items as well as items with IRFs that are not well approximated by any three-parameter logistic IRF. At this point, flawed, overly difficult, and otherwise atypical items can be revised or deleted and the researcher can decide whether the characteristics of the non-three-parameter logistic items justify continued use of the nonparametric model (e.g., the items might be too discriminating in some θ range to be modeled with three-parameter logistic ogives). If departures from three-parameter logistic ogives are not too severe, the test developer then could use parametric methods such as BILOG or LOGIST 5 for subsequent analyses.

Appendix: MFS Test Calibration

BILOG was used to initialize the maximum likelihood formula scoring programs in the following way. First, IRFs estimated by BILOG were used to compute basis functions for density estimation. The θ density was represented as a linear combination of basis functions with unknown coefficients. The probability of each item response pattern then was expressed as a linear function of these unknown coefficients. The sample likelihood function was expressed as the product of these linear functions. A maximum likelihood density estimate was obtained by maximizing the logarithm of the sample likelihood function subject to certain constraints. The estimated density was constrained to be nonnegative, increasing between −3 and −1, and decreasing between 1 and 3. 61 evenly spaced (between −3 and 3) quadrature points were used for the integrations.

ForScore (Williams & Levine, in preparation) then used the estimated θ density and marginal maximum likelihood estimation to determine the coordinates of the IRFs with respect to the orthonormal basis. Thus, each IRF was represented as a linear combination of basis functions, and the coefficients of the linear combination were estimated by marginal maximum likelihood estimation. All IRFs were estimated, one at a time, in each of seven cycles. During a cycle, the likelihood function for each IRF was evaluated using the provisional estimates of the other IRFs. For each item, 13 coordinates were estimated using 61 quadrature points. On the first cycle, IRFs were constrained to be between .005 and .999, and the absolute values of their third derivatives were constrained to be ≤ 1. On the remaining cycles, the constraint on the third derivative was relaxed for rapidly rising segments of IRFs. A demonstration of the use of MFS was provided by Drasgow, Levine, Williams, McLaughlin, & Candell (1989).

References

Bickel, P. J., & Docksum, K. A. (1977). *Mathematical statistics: Basic ideas and selected topics.* San Francisco: Holden-Day.

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

Candell, G. L. (1988). Application of appropriateness measurement to a problem in adaptive testing (Doctoral dissertation, University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 50, (part 2b),* 782.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129–141.

Davey, T., & Hirsch, T. (1991, April). *Examinee discrimination and the measurement properties of multidimensional tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13,* 77–90.

Drasgow, F., Levine, M. V., Williams, B., McLaughlin, M. E., & Candell, G. L. (1989). Modelling incorrect responses to multiple-choice items with multilinear formula score theory. *Applied Psychological Measurement, 13,* 285–299.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

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

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics* (Vol. 2, 4th ed.). New York: Macmillan.

Lazarsfeld, P. F. (1959). Latent structure analysis. In S. Koch (Ed.), *Psychology: A study of a science* (Vol. 3) (pp. 476–542). New York: McGraw-Hill.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* New York: Wiley.

Levine, M. V. (1984). *An introduction to multilinear formula score theory* (Measurement Series 84–4). Champaign: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V. (1989). *Classifying and representing ability distributions* (Measurement Series 89-1). Champaign: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (1992). *Measuring the difference between two models* (Measurement Series 92-1). Champaign: University of Illinois, Department of Educational Psychology, Model-Based Measurement Laboratory.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4,* 269–290.

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

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 51–61). New York: Academic Press.

Mislevy, R. J., & Bock, R. D. (1984a). *BILOG II user's guide* [Computer program manual]. Mooresville IN: Scientific Software, Inc.

Mislevy, R. J., & Bock, R. D. (1984b). *Item operating characteristics of the Armed Services Vocational Aptitude Battery (ASVAB), Form 8A.* Unpublished manuscript.

Samejima, F. (1979). *A new family of models for multiple choice items* (Research Rep. No. 79-4). Knoxville: University of Tennessee, Department of Psychology.

Sympson, J. B. (1986, April). *Some item response functions obtained in polychotomous item analysis.* Paper presented at the Office of Naval Research Contractors' Meeting on Model-Based Psychological

Measurement, Gatlinburg TN.

Sympson, J. B. (1988, May). *A procedure for linear polychotomous scoring of test items.* Paper presented at the Office of Naval Research Contractors' Meeting on Model-Based Psychological Measurement, Iowa City IA.

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

Williams, B., & Levine, M. V. (in preparation). *ForScore: A computer program for nonparametric item response theory.*

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide* [Computer program manual]. Princeton NJ: Educational Testing Service.

Wood, R. L., Wingersky, M. S., & Lord, F. M. (1976). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (Res. Memo. 76-6). Princeton NJ: Educational Testing Service.

Acknowledgments

The authors thank Daniel O. Segall for his help with many aspects of the research reported here; J. Bradford Sympson for providing item parameter estimates for Studies 3 and 4 and for sharing computer code for his polyweight model; and Robert J. Mislevy, Martha L. Stocking, and Marilyn S. Wingersky for their comments on an earlier draft of this paper and for conducting the LOGIST 5 *analyses. This research was supported primarily by the Naval Personnel Research and Development Center. Michael Levine's participation was supported by Office of Naval Research Contract No. N00014-86K-0482, NR 442-1546.*

Author's Address

Send requests for reprints or further information to Michael V. Levine, 210 Education Building, University of Illinois, 1310 South Sixth Street, Champaign IL 61820, U.S.A.