Fitting Polytomous Item Response Theory Models to Multiple-Choice Tests

Fritz Drasgow, Michael V. Levine, Sherman Tsien, Bruce Williams, and Alan D. Mead

University of Illinois

This study examined how well current software implementations of four polytomous item response theory models fit several multiple-choice tests. The models were Bock's (1972) nominal model, Samejima's (1979) multiple-choice Model C, Thissen & Steinberg's (1984) multiple-choice model, and Levine's (1993) maximum-likelihood formula scoring model. The parameters of the first three of these models were estimated with Thissen's (1986) MULTILOG computer program; Williams & Levine's (1993) FORSCORE program was used for Levine's model. Tests from the Armed Services Vocational Aptitude Battery, the Scholastic Aptitude Test, and the American College Test Assessment were analyzed. The models were fit in estimation samples of approximately 3,000; cross-validation samples of approximately 3,000 were used to evaluate goodness of fit. Both fit plots and χ^2 statistics were used to determine the adequacy of fit. Bock's model provided surprisingly good fit; adding parameters to the nominal model did not yield improvements in fit. FORSCORE provided generally good fit for Levine's nonparametric model across all tests. *Index terms: Bock's nominal model, FORSCORE, maximum likelihood formula scoring, MULTILOG, polytomous IRT.*

Since the introduction of the Bock (1972) nominal model (BNM), polytomous item response theory (IRT) models have been viewed as a means for improving psychological measurement. For example, in comparison to dichotomous models, polytomous models allow more information about trait level to be extracted from a fixed set of items (Bock, 1972; Drasgow, Levine, Williams, McLaughlin, & Candell, 1989; Sympson, 1986, 1993; Thissen, 1976; Thissen & Steinberg, 1984), provide increased rates of detection of aberrant response patterns (Drasgow, Levine, & McLaughlin, 1987, 1991), and can be used to provide specific feedback to item writers about which distractors are effective and which are ineffective. Such benefits, however, may not be realized if a polytomous model inadequately fits a dataset. In the research described here, a series of models of increasing complexity was applied to data from several multiple-choice tests to examine the degree of generality needed by a polytomous model to fit item responses adequately.

The models examined here include the BNM, the Samejima (1979) multiple-choice Model C (SMCM), the Thissen & Steinberg (1984) multiple-choice model (TSMCM), and the Levine (1993) polytomous maximum-likelihood formula scoring model (MFSM). The BNM, the SMCM, and the TSMCM are nested parametric models in the sense that the simpler models can be obtained from the more complex models by setting some parameters to 0. The MFSM is the most general of the models studied because its option response functions (ORFs) do not have a specific parametric form; because it is a nonparametric model, its ORFs may assume a much wider variety of shapes. This study examined the degree to which the nonparametric model's greater flexibility produced improved fit in cross-validation samples.

There are polytomous IRT models that are more restrictive than any of the models examined here (e.g., Andrich, 1978; Masters, 1982; Thissen & Steinberg, 1986) and methods that do not use the logistic model combined with maximum likelihood estimation [e.g., Sympson's (1986) Model 8 and polyweight analysis

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 19, No. 2, June 1995, pp. 143–165
© Copyright 1995 Applied Psychological Measurement Inc.
0146-6216/95/020143-23$2.40

(1988), and Samejima's (1983) simple sum and differential weights models]. The more restrictive models were not included in this study because preliminary analyses indicated lack of fit. Abrahamowicz & Ramsay's (1992) multicategory spline model would have been included in this study if their software had been available at the time the research was conducted.

Polytomous Models

Let **v** = [v_1, v_2, ..., v_i, ..., v_n] denote the random vector of polytomously scored responses to n items, and let **v*** = [v_1*, v_2*, ..., v_i*, ..., v_n*] denote a specific response pattern. For multiple-choice items with s options (or categories), v_i is scored 1 if the first option is selected by an examinee, 2 if the second option is selected, ..., k if the kth option is selected, ..., and s if the last option is selected. Assume that the items have been recoded so that the first option is always the correct option.

All of the models considered here are unidimensional models. Let θ denote the latent trait with density $f(\cdot)$. A specific value of θ is denoted t.

Bock's Nominal Model

In this model, the probability of selecting option k on item i is written

$$P(v_i = k|\theta = t) = \frac{\exp(a_{ik}t + c_{ik})}{\sum_{k'=1}^{s}\exp(a_{ik'}t + c_{ik'})}. \tag{1}$$

In Bock's (1972) approach to maximum likelihood estimation of item parameters, the a_{ik} and the c_{ik} are parameters associated with the kth option of item i and are constrained to sum to 0 for each item i. These constraints also are imposed in MULTILOG (Thissen, 1986).

Assuming that each option has a distinct a_{ik} value, the option with the largest a_{ik} will have an ORF that monotonically increases to 1.0, so that individuals with sufficiently high θs will select this option with high probability. Ordinarily Option 1, the correct option, would be expected to have the largest a_{ik} and so its ORF would be monotonically increasing. The option with the smallest a_{ik} (i.e., the largest negative a_{ik}) will have an ORF that monotonically decreases from a left tail value of 1.0 to a right tail value of 0. Thus, according to the BNM, almost all examinees at low θ levels will select the same incorrect option (i.e., the option with the smallest a_{ik}; see Samejima, 1972). Empirical research shows that multiple-choice tests such as the Scholastic Aptitude Test (SAT), the Armed Services Vocational Aptitude Battery (ASVAB), and the American College Test (ACT) do not have this property (Levine & Drasgow, 1983).

Samejima's Multiple-Choice Model

In an attempt to allow nonzero left tails for all ORFs, Samejima (1979) introduced the idea of a latent response category, which Thissen & Steinberg (1984) called the latent "don't know" (DK) category. One way to view the DK category is that it characterizes examinees who do not have any idea of the answer to an item. This psychological state is different than a response state in which an examinee knows the answer or falsely believes that one of the incorrect options is the answer. Thus, examinees in the DK state, if forced to select an answer, will simply guess.

The SMCM incorporates a conditional probability that a randomly sampled examinee at each θ value will fall in the DK category. A plot of these probabilities might be considered to be the ORF for the latent DK category. On tests with number-correct scoring, however, examinees in the DK state will ordinarily select one of the response options (and even on tests with corrections for guessing many examinees evidently feel compelled to answer when they are uncertain). Thus, the probability from the DK ORF is distributed to the ORFs for the observed response categories.

In the SMCM, the probability of the DK state is

$$P(\mathrm{DK}|\theta = t) = \frac{\exp(a_{i0}t + c_{i0})}{\sum_{k'=0}^{s} \exp(a_{ik'}t + c_{ik'})}, \tag{2}$$

where a_{i0} and c_{i0} are the parameters for the DK state. Samejima assumed that DK examinees randomly select one of the s manifest response options. Specifically, each manifest response category is selected by a DK examinee with probability $d_{ik} = 1/s$. This assumption leads to

$$P(v_i = k|\theta = t) = \frac{\exp(a_{ik}t + c_{ik}) + d_{ik}\exp(a_{i0}t + c_{i0})}{\sum_{k'=0}^{s} \exp(a_{ik'}t + c_{ik'})}. \tag{3}$$

In sum, the SMCM is a generalization of the BNM that incorporates additional parameters designed to allow the left tails of all ORFs to have asymptotes between 0 and 1. This was accomplished by adding two additional parameters (a_{i0} and c_{i0}) per item. Note that if all the a_{ik} are distinct and if a_{i0} is the smallest a parameter for item i, then the left tails of all ORFs will asymptote at $1/s$. If a_{i0} is not the smallest a, then the ORF for the option with the smallest a will have a left tail that asymptotes at 1 and all other ORFs will have left tails that go to 0.

Thissen and Steinberg's Multiple-Choice Model

Thissen & Steinberg (1984) found the assumption that DK examinees guess at random to be implausible. Instead, they believed that different options might attract DK examinees at differential rates. Thissen and Steinberg used this idea to generalize the SMCM by treating the d_{ik} as parameters to be estimated, rather than fixed constants. Consequently, the mathematical form of the TSMCM is identical to the SMCM; the difference lies in the way that the d_{ik} are treated.

Thissen & Steinberg's (1984) main contribution was the development of a method for estimating the d_{ik} in addition to the a_{i0} and c_{i0}. This culminated in their **T** matrix formulation of polytomous models (Thissen & Steinberg, 1986), which serves as the theoretical foundation of MULTILOG (Thissen, 1986).

Levine's Maximum Likelihood Formula Scoring Model

The previous models have several common features. They utilize functions with a small number of parameters, which are used in nonlinear formulas to define ORFs. The likelihood of the sampled data can be written as an explicit function of the parameters, which thereby allows maximum likelihood estimation.

The basic structure of the MFSM can be developed by reference to these features. The MFSM uses a larger number of parameters that combine linearly to represent ORFs with arbitrary shapes. Nonetheless, the likelihood of the sampled data can be written as an explicit function of the parameters, and the parameters and the ORFs can be estimated using maximum likelihood estimation.

The MFSM represents functions as linear combinations of a finite set of orthogonal functions. Some well-known sets of orthogonal functions include orthogonal polynomials and trigonometric functions. The shape of ORF estimates obtained by the MFSM does not depend on which orthogonal functions are used.

The basic formula for the MFSM is

$$P(v_i = k|\theta = t) = \sum_j \alpha_{ijk} h_j(t). \tag{4}$$

Here the h_j are the orthogonal functions and the αs are numerical constants to be estimated by marginal maximum likelihood estimation. The summation index j ranges from 0 to J where J in this application was 8, which Williams (1986) found to be satisfactory for dichotomously scored items.

If every item except item *i* has been modeled, then the conditional likelihood of response *k* to item *i* can be written in the linear form

$$P\left(v_i = k, \text{ pattern } \mathbf{v}^* \text{ on the remaining } n-1 \text{ items} \mid \theta = t\right) = \sum_j \alpha_{ijk} h_j(t) l(\mathbf{v}^*, t). \tag{5}$$

where $l(\mathbf{v}^*, t)$ denotes the likelihood of $\mathbf{v}^*$ (without item *i*) at *t*. Thus, the marginal likelihood of the *n*-item response pattern is

$$\sum_j \alpha_{ijk} \int h_j(t) l(\mathbf{v}^*, t) f(t)\, dt, \tag{6}$$

where *f* is the density of θ, so that $l(\mathbf{v}^*, t)f(t)$ is proportional to the posterior θ density given $n - 1$ item responses.

The MFSM makes a simplifying assumption at this point. Because the total number of response patterns, although very large, is finite, the set of posterior densities can be written as linear combinations of a finite number of functions. The MFSM assumes that the vector spaces of functions obtained as linear combinations of posterior densities of $n - 1$ item patterns are very nearly the same no matter which item is left out. This implies that the ORFs for the left out item can be closely approximated as a linear combination of the posterior densities computed from the remaining items. For this reason, the MFSM takes as its h_js linear combinations of posterior densities computed from its current model for the test.

To keep the number of h_j tolerably small, the MFSM selects a set of *J* orthogonal functions that can approximate the entire array of posterior densities with the smallest total mean squared error. Note that if the posterior distributions of the provisional model used to compute the h_js have several continuous derivatives, then the h_js also will have several continuous derivatives.

In typical parametric theories, the possible values of the parameters to be estimated are restricted. For example, in the three-parameter logistic model, the lower asymptote parameter is constrained to be between 0 and 1, the item discrimination parameter is constrained to be positive, and item difficulty is typically constrained to lie between −3 and +3 to avoid implausible values.

The MFSM is implemented in the computer program FORSCORE (Williams & Levine, 1993), which also uses constrained optimization. FORSCORE is designed to translate qualitative assumptions about the shapes of functions into linear equalities that must be satisfied during the optimization process. Thus, the user can impose constraints so that the ORFs for correct options are monotone increasing and find the best-fitting monotonic model. The condition that an ORF is nondecreasing at *t* is simply

$$\frac{d}{dt}P\left(v_i = k \middle| \theta = t\right) = \sum_j \alpha_{ijk} \frac{d}{dt} h_j(t) = \sum_j \alpha_{ijk} h_j'(t) \ge 0. \tag{7}$$

Monotonicity can be assumed globally or just over an interval, for some or for all ORFs.

Two additional types of constraints implemented in FORSCORE are concavity and smoothness. By using higher derivatives, the user can constrain some of the estimated functions to be concave or convex over specified intervals. Finally, FORSCORE permits constraints that force ORFs to be smooth. By constraining the values of the first, second, or third derivatives, the user can force estimated ORFs to be approximately linear, quadratic, or cubic, respectively, over an interval.

Goodness of Fit

Assessing the fit of different models is difficult. Adherence to the significance of a test of fit is inappropriate because all IRT models are misspecified to some degree and, therefore, a significance test in a very large sample will almost certainly reject any IRT model. Consequently, a combination of complementary graphical and statistical methods were used to evaluate fit.

Fit Plots

Earlier work by Drasgow et al. (1989) often found conditional ORFs (CORFs) to be more informative than ORFs. A CORF provides the probability of selecting an incorrect option as a function of θ in the subpopulation of examinees who answered an item incorrectly. An item's CORF is therefore related to its ORF by CORF = ORF/[1 − $P_i(\theta)$]. This relation shows why CORFs are useful: because ORF = [1 − $P_i(\theta)$] × CORF, the ORFs for incorrect options are compressed toward 0 as $P_i(\theta)$ increases to 1. However, CORFs sum to 1 at all θ values and can thus provide information about incorrect options that are differentially attractive to examinees in different θ ranges.

The process of constructing a fit plot begins by estimating item response functions (IRFs) and ORFs in a test calibration sample. MULTILOG explicitly assumes that these response functions are stated in reference to a θ distribution that is standard normal. A FORSCORE option was used to scale the MFSM response functions so that they too were estimated in reference to a standard normal θ distribution.

A cross-validation sample disjoint from the sample used to estimate ORFs was used to determine empirical proportions of option selection. A cross-validation sample was used because it is important to avoid artifacts that result from overfitting when contrasting models with varying numbers of parameters. Models with greater numbers of parameters may simply reflect sampling fluctuations in the test calibration sample; thus, a cross-validation sample was used to evaluate fit in order to be fair to the simpler models.

In usual approaches to fit plots, the θ continuum is divided into, say, 25 strata. θ is estimated for an examinee, and then the total number of examinees in each θ stratum is counted as well as the number of examinees who selected each option. An empirical proportion is computed as the number of examinees who selected option *k* divided by the total number of examinees in the stratum.

The problem with this straightforward approach to constructing fit plots is that the θ estimate ($\hat{\theta}$) is not ordinarily equal to θ, and the degree of error in $\hat{\theta}$ is ignored. Error in $\hat{\theta}$ may cause the fit plot to be smoother than the response function it estimates. Bias in $\hat{\theta}$s changes the shape of the fit plot in complex ways. Thus, even with a very large sample and perfectly estimated response functions, the sample fit plot may differ substantially and systematically from the true response function. In fact, because the shape of the expected fit plot depends on the particular θ estimator in its construction, one set of fit plots may favor one model and another set of fit plots computed with a different θ estimator may favor a different model.

A simple solution to these problems can be found in Levine & Williams's (1991, 1993) extension of Samejima's (1983) simple sum procedure. Levine and Williams demonstrated that even for a biased estimator with substantial sampling error, appropriately constructed fit plots can have the same shape as the true IRF.

In Samejima's model, the simple sum estimate of a point $P_i(t)$ on an IRF is computed with an estimate $\hat{\tau}$ of ability. Specifically, it has the form,

$$\hat{P}_i(t) = \frac{\sum\limits_{\{a:a\in S^+\}} P\{\theta = t \mid \hat{\tau} = \hat{\tau}_A\}}{\sum\limits_{A} P\{\theta = t \mid \hat{\tau} = \hat{\tau}_A\}} . \qquad (8)$$

In Equation 8, the summation in the denominator is over all examinees in the sample, and the summation in the numerator is over only the examinees who correctly answered the item. $\hat{\tau}$ is a θ estimator computed from responses to all items except the target item, item *i*, and $\hat{\tau}_A$ is the value of $\hat{\tau}$ computed from Examinee A's data.

Of course, some $\hat{\tau}$ statistics will make better use of the item responses and provide more useful fit plots. Similarly, the conditional distributions in the simple sum formula will be more easily computed for some statistics. Levine & Williams (1991, 1993) proposed replacing the statistic $\hat{\tau}$ with the vector-valued statistic giving the examinee's responses to all but the target item response.

Thus, using the dichotomously scored item response pattern $\mathbf{u}^*$ for $\hat{\tau}$, Levine & Williams' (1991, 1993) extension of the simple sum formula can be written

$$\hat{P}_i(t) = \frac{N^+}{N} \frac{\sum_{\{a: a \in S^+\}} P\{\theta = t | \mathbf{u} = \mathbf{u}^*_{\mathrm{A}}\}/N^+}{\sum_{\mathrm{A}} P\{\theta = t | \mathbf{u} = \mathbf{u}^*_{\mathrm{A}}\}/N}, \tag{9}$$

where N^+ is the number of examinees correctly answering the target item (or, in polytomous applications, N^+ would be the number of examinees selecting a specified option), and N is the total number of examinees. Note that the fit plot is proportional to the ratio of two averaged posterior θ densities: The denominator is averaged over all examinees and the numerator is averaged over a subsample. This is the equation that would be obtained if—rather than assigning an examinee to a cell and incrementing the count in this cell, as in the usual fit plot histogram—the examinee's posterior density was used to distribute the "count" over the θ continuum.

Using only smoothness conditions that are valid for all the models considered here, Levine's $\hat{P}_i(t)$ is a strongly consistent estimator of $P_i(t)$ (Levine, 1988, 1989; Levine & Williams, 1991, 1993). In other words, $\hat{P}_i(t)$ will approach a point on the true IRF as the sample size is increased with probability 1.0, provided, of course, that the response functions are correctly specified. Moreover, Levine & Williams' (1991, 1993) use of the response vector in the above formula gives the fit plot an easily computed, intuitively appealing interpretation. These plots are similar to those reported by Mislevy & Bock (1989) who used posterior θ densities in their evaluation of IRF estimates obtained with BILOG (Mislevy and Bock).

In sum, a straightforward generalization of Equation 9 for polytomous item responses was used to calculate fit plots as ratios of averaged posterior densities. The fit plots were constructed using a grid of 25 points, with grid points selected as the 2nd, 6th, ..., 98th percentile points from the standard normal distribution.

χ^2 Fit Statistics

As with the fit plots, the χ^2 approach began by taking a cross-validation sample of N examinees. The same cross-validation sample was used to construct fit plots and to compute the χ^2 statistics.

There are n χ^2 statistics that can be computed for the n items individually. However, there are $\binom{n}{2}$ χ^2 statistics that can be computed for item pairs and $\binom{n}{3}$ possible χ^2 statistics for item triples. To limit the number of χ^2 statistics to a manageable (and comprehensible) number, the n test items were divided into $n/3$ sets of three items. For each set, a χ^2 was computed for each item, for all three sets of item pairs, and for the triple of items. The sets were selected so that they each contained a relatively easy item, an item of moderate difficulty, and a relatively difficult item.

To compute the χ^2 for item i, the expected number of times that examinees would select option k was computed from the ORF using

$$E_i(k) = N \int P(v_i = k | \theta = t) f(t)\, dt, \tag{10}$$

where $f(\cdot)$ is the θ density, taken to be the standard normal because ORFs were scaled in reference to this distribution. The above integral was evaluated by numerical quadrature using 161 grid points on the interval [−3,+3]. The observed frequency $O_i(k)$ of option k was determined by simply counting the number of times examinees selected this option in the cross-validation sample. Finally, the ordinary χ^2 for item i was computed from the expected and observed frequencies,

$$\chi_i^2 = \sum_{k=1}^{s} \frac{[O_i(k) - E_i(k)]^2}{E_i(k)}. \tag{11}$$

To compute the χ^2 for items i and m, the expected frequency of the (k, k')th cell in the two-way table was computed by

$$E_{l,m}(k,k') = N\int P(v_i = k|\theta = t)P(v_m = k'|\theta = t)f(t)\,dt, \tag{12}$$

and observed frequencies for the two-way table were counted. Cells with expected frequencies less than 5 were aggregated; if the sum of these expected frequencies was still less than 5, the cells were combined with the cell with expected frequency that least exceeded 5. Finally, the usual χ^2 for a two-way table was calculated. The χ^2 statistic for each item triple was computed by the obvious extension of the above procedure.

Method

Tests

Data from three test batteries were analyzed by the four polytomous models. These test batteries included the ASVAB, the SAT, and the ACT.

ASVAB. The first dataset contained the responses of 13,569 examinees who completed the ASVAB, Form 17A (Department of Defense, 1984), under operational conditions (i.e., they were administered the ASVAB as part of the standard military enlistment process). Four ASVAB subtests were used in the analyses. First, the 30-item Arithmetic Reasoning subtest and the 25-item Math Knowledge subtest were combined to form a 55-item quantitative reasoning test. One item was deleted because it was too easy for meaningful analyses, leaving a total of 54 items. The 35-item Word Knowledge subtest and the 15-item Paragraph Comprehension subtest also were combined to form a 50-item verbal test.

The test calibration sample consisted of a sample of 3,392 examinees. This sample was formed by taking every fourth examinee, beginning with the first examinee. Thus, examinees 1, 5, 9, ... were included. [A simple notation for this sample is (4, 1): Take every fourth examinee beginning with examinee number 1.] The ASVAB cross-validation sample consisted of 3,392 examinees obtained with the sampling plan (4, 2).

SAT. The second dataset consisted of the responses of approximately 108,000 examinees from the November 1989 administration of the SAT (Donlon, 1984). Both the 85-item SAT verbal (SATV) test and the 60-item SAT math (SATM) test were analyzed. The SAT test calibration sample consisted of 3,000 examinees obtained with the sampling plan (36, 1). The cross-validation sample contained 3,000 examinees obtained from a (36, 2) sampling plan.

ACT. The responses of 140,979 examinees from the October 1983 administration of the ACT Assessment (American College Testing Program, 1988) constituted the third dataset. The 40-item Math Usage test was used for these analyses. The sampling plan (41, 1) was used to form the ACT test calibration sample of N = 3,000 examinees. The cross-validation sample, also consisting of 3,000 response patterns, was obtained using a (41, 2) sampling plan.

Analysis

Implementing FORSCORE. In implementing FORSCORE, only global smoothness constraints were applied to the ORFs so that the large dataset would determine the shapes of ORFs. Third derivatives were globally constrained to be small in absolute value. (Of course, probabilities were constrained to be positive and sum to 1; however, no other shape-controlling constraints were applied.)

The FORSCORE implementation of the MFSM represents ORFs as linear combinations of specified smooth functions, the h_js. The coefficients in the linear combination are estimated in cycles, one item at a time, to increase the likelihood of a fixed sample of polytomous data. Linear equations and equalities constraining the coefficients of linear combination force the estimated functions to be positive, to sum to 1, and to have other qualitative features specified by the user. The estimated model is iteratively revised until no further increase in likelihood is observed.

This study departed from this general implementation of FORSCORE in two ways. First, to speed convergence, FORSCORE was permitted to aggregate the incorrect options and analyze the data as if it were dichotomous to obtain the ORFs for correct responses. After the dichotomous analysis converged, the data were analyzed polytomously. In the successive updates of the polytomous MFSM, FORSCORE was not permitted to revise its estimate of the ORFs for correct options.

Finally, in order to avoid dealing with very small numbers, CORFs were used for incorrect options. Thus, response functions for the incorrect options were represented as

$$P(v_i = k|\theta = t) = \left[1 - P(v_i = 1|\theta = t)\right] \cdot P^*(v_i = k|\theta = t), \quad k = 2, 3, \ldots, s, \qquad (13)$$

where the αs in

$$P^*(v_i = k|\theta = t) \; = \sum_j \alpha_{ijk} h_j(t), \quad k = 2, 3, \ldots, s, \qquad (14)$$

were estimated. As is the case with the αs for the correct options, the MFSM used marginal maximum likelihood to estimate the α_{kij}.

Test calibration. The parameters of the BNM, the SMCM, and the TSMCM were estimated with Thissen's (1986) MULTILOG computer program; Williams & Levine's (1993) FORSCORE program was used for the MFSM. An attempt was made to analyze the tests from the three test batteries consistently. Initially, omitted responses were excluded from the likelihood function. Unfortunately, MULTILOG failed to converge properly for the SAT tests. Because the SAT penalizes examinees for incorrect responses, many examinees omitted a substantial number of items and this created problems for MULTILOG. Therefore, omits were treated as a response category for all three test batteries and ORFs were estimated. These analyses, however, were unsuccessful for the ASVAB and ACT because MULTILOG provided many extreme item parameter estimates (e.g., −13 and +9). The ASVAB and the ACT use number-correct scoring; consequently, there was very little omitting and it was not surprising that MULTILOG encountered difficulties in analyses requiring estimation of an ORF for this category.

The most reasonable results were obtained with MULTILOG when an IRT model was used that was consistent with the directions given to examinees. Therefore, omits were excluded from the likelihood function for the two test batteries that used number-correct scoring and had very little omitting (the ASVAB and the ACT). However, omit was treated as a response category for the SAT, which instructs examinees to omit when they do not know the answer and therefore has many omitted responses. FORSCORE, however, provided reasonable ORF estimates for all three test batteries, both when omits were treated as a response category and when omits were excluded from the likelihood function.

Another problem was encountered in the analyses using MULTILOG. Depending on the complexity of the model, only 20 to 30 items could be calibrated simultaneously, which evidently is a consequence of the 640K memory limitation imposed by DOS (the DOS version of MULTILOG was used). Unfortunately, the results provided by MULTILOG for the SMCM and for the TSMCM sometimes depended on which set of 20 to 30 items was analyzed. For example, the ORF estimated for the correct option of a particular item might be monotonically increasing when it was analyzed with one set of items (i.e., this is what would be expected for a correct response) but nonmonotonic (first increasing, then decreasing for moderate and high θ values) when analyzed with another set of items. This problem was particularly evident for the SAT: MULTILOG sometimes provided monotonically increasing ORFs for the omit option and nonmonotonic ORFs for the correct option for all items.

The results presented here represent the outcome of analyzing and reanalyzing the datasets with MULTILOG until "reasonable" ORFs (i.e., monotonically increasing ORFs for the correct options) were obtained for all items in a test. No reanalysis was needed for the BNM. However, substantial problems were encountered with the two more general models (the SMCM and the TSMCM) and many analyses were required.

FORSCORE, which runs under the Unix operating system, can analyze up to 100 items in a single run. Thus, all items of each test were analyzed simultaneously.

Cross-validation. After calibrating a test, the fit of the model was evaluated with the cross-validation sample. To summarize the large number of χ^2 statistics, ratios of χ^2s to their degrees of freedom (*df*) were computed and tabulated for four intervals: very small (<1), small (≥1 and <2), moderately large (≥ 2 and < 3), and large (≥ 3). Fit plots were constructed for each item with the appropriate cross-validation sample.

Results

ASVAB

Table 1 contains the χ^2/df ratios for the ASVAB quantitative and verbal tests. Surprisingly, the more flexible SMCM and TSMCM showed little or no improvement in fit over the simpler BNM. The MFSM, however, had noticeably smaller χ^2/df statistics for the quantitative test. For example, 36 of the 54 items on this test had χ^2/df in the very small or small ranges; the three parametric models had 28, 23, and 20 χ^2/df in the very small (<1) or small (1–<2) range. The mean χ^2/df for single items was 1.81 for the MFSM; the means were 2.46, 2.93, and 2.98 for the BNM, SMCM, and TSMCM, respectively. For χ^2s computed for pairs of items, the MFSM had 40 χ^2/df of less than 2; the three parametric models had 23, 23, and 19 ratios in these ranges. Again the mean of the ratios was considerably smaller for the MFSM: 1.80 versus 2.41, 2.46, and 2.44 for the three parametric models. Finally, the χ^2/df statistics for item triples also indicated a better fit for the MFSM.

The fit of the three parametric models and, to a lesser extent, the MFSM, were somewhat better for the ASVAB verbal test than for the quantitative test. (An explanation for the improved fit is suggested by the fit plots, and is discussed below.) The fit of all four models to the verbal test was relatively similar. The BNM had the smallest mean χ^2/df (1.81) for single items; the MFSM had the smallest means for χ^2/df computed for pairs of items (1.44) and triples of items (1.24).

Space limitations prohibit an extended presentation of the fit plots: 216 plots were constructed for the quantitative test (one plot for each of the 54 items as analyzed by each of the four models) and 200 plots were constructed for the verbal test. Figure 1, which shows Item 19 from the quantitative ASVAB, shows a typical fit plot for the BNM. The responses for Figures 1–7 were recoded so that Option 1 was always the correct option, and the response functions for the incorrect options (Options 2, 3, 4, etc.) were CORFs $\left[P^*\left(v_i = k|\theta = t\right)\right]$ that sum to unity at each θ value:

$$P^*\left(v_i = 2|\theta = t\right) + P^*\left(v_i = 3|\theta = t\right) + P^*\left(v_i = 4|\theta = t\right) = 1, \text{ for all } t. \tag{15}$$

The vertical lines in Figure 1 depict approximate 95% confidence intervals for the empirical pseudo-proportions. The confidence intervals for the correct option (Option 1) were computed as

$$\hat{P}_i(t) \pm 2\left\{\frac{\hat{P}_i(t)\left[1-\hat{P}_i(t)\right]}{N^*}\right\}^{½}, \tag{16}$$

where $\hat{P}_i(t)$ was defined in Equation 1, and N^* is the sum of the posterior densities evaluated at $\theta = t$. Confidence intervals for the CORFs were computed analogously. A confidence interval was not plotted at $\theta = t$ if the sum of the posterior densities for that response function was less than 5.

The empirical proportions in Figure 1 indicate that a nonzero lower asymptote would be appropriate for the correct option (Figure 1a). The functional form of the BNM forces the estimated IRF to a lower asymptote of 0, and thus it appears that the BNM seeks to estimate a response function that is too complex for its mathematical form. The systematic error in Figure 1 resulted in a moderately large χ^2/df of 3.08.

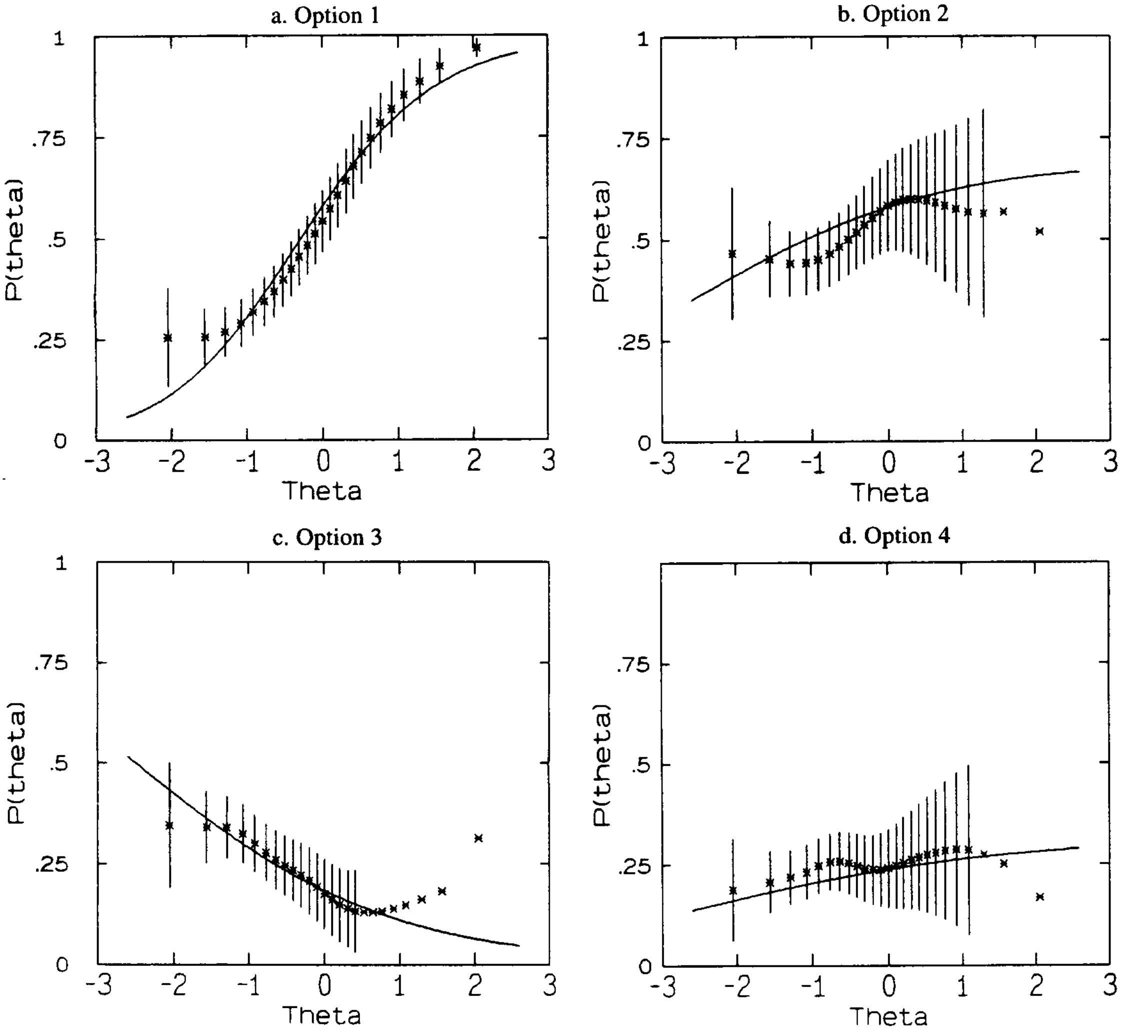
Table 1
Frequencies, Means, and Standard Deviations (SDs) of ASVAB χ^2/df Ratios

Test, Model,	Frequency of χ^2/df					
and Items	<1	1–<2	2–<3	≥3	Mean	SD
Quantitative Test						
BNM						
Singles	15	13	8	18	2.46	2.17
Doubles	1	22	21	10	2.41	.97
Triples	0	5	11	2	2.22	.59
SMCM						
Singles	15	8	15	16	2.93	3.41
Doubles	4	19	19	12	2.46	1.45
Triples	0	12	5	1	2.02	.59
TSMCM						
Singles	11	9	10	24	2.98	2.01
Doubles	1	18	23	12	2.44	1.04
Triples	0	13	5	0	1.94	.48
MFSM						
Singles	18	18	10	8	1.81	1.39
Doubles	5	35	10	4	1.80	1.04
Triples	1	14	2	1	1.63	.51
Verbal Test						
BNM						
Singles	19	13	7	11	1.81	1.48
Doubles	15	23	11	0	1.62	.61
Triples	1	13	2	0	1.48	.33
SMCM						
Singles	16	13	9	12	2.56	3.48
Doubles	8	27	9	5	1.78	1.03
Triples	2	12	2	0	1.48	.47
TSMCM						
Singles	18	15	9	8	2.04	2.00
Doubles	8	32	5	4	1.54	.72
Triples	4	10	2	0	1.32	.40
MFSM						
Singles	19	13	6	12	1.96	1.71
Doubles	11	31	6	1	1.44	.64
Triples	3	13	0	0	1.24	.32

The CORFs shown in Figure 1 for the three incorrect options (Figures 1b–1d) were typical for the BNM. Repeatedly, these functions were found to be nearly linear, which was both a benefit and a disadvantage. The advantage was that across hundreds of fit plots substantial misfit of the BNM to empirical proportions was rarely observed. The disadvantage was that the BNM frequently showed systematic (but relatively small) errors that a more complex model should remedy.

Figures 2 and 3 present CORFs estimated by the TSMCM for two quantitative items (Items 52 and 53, respectively). The estimated functions for Item 53 are much more satisfactory than the functions estimated for Item 52 because they better fit the empirical proportions. Interestingly, the χ^2s suggest a slight superiority for Item 52 (χ^2 = 10.36 for Item 52 versus 12.34 for Item 53). The relatively large χ^2 for Item 53 was due mainly to the correct option (Figure 3a), because the estimated function was below the empirical proportions in almost every θ stratum. The observed number of examinees selecting the correct response was 1,097 and the number expected given the estimated response function was 1,008.41. The difference, 88.59, when squared

Figure 1
ASVAB Quantitative Item 19 Analyzed by the BNM


a. Option 1

b. Option 2

c. Option 3

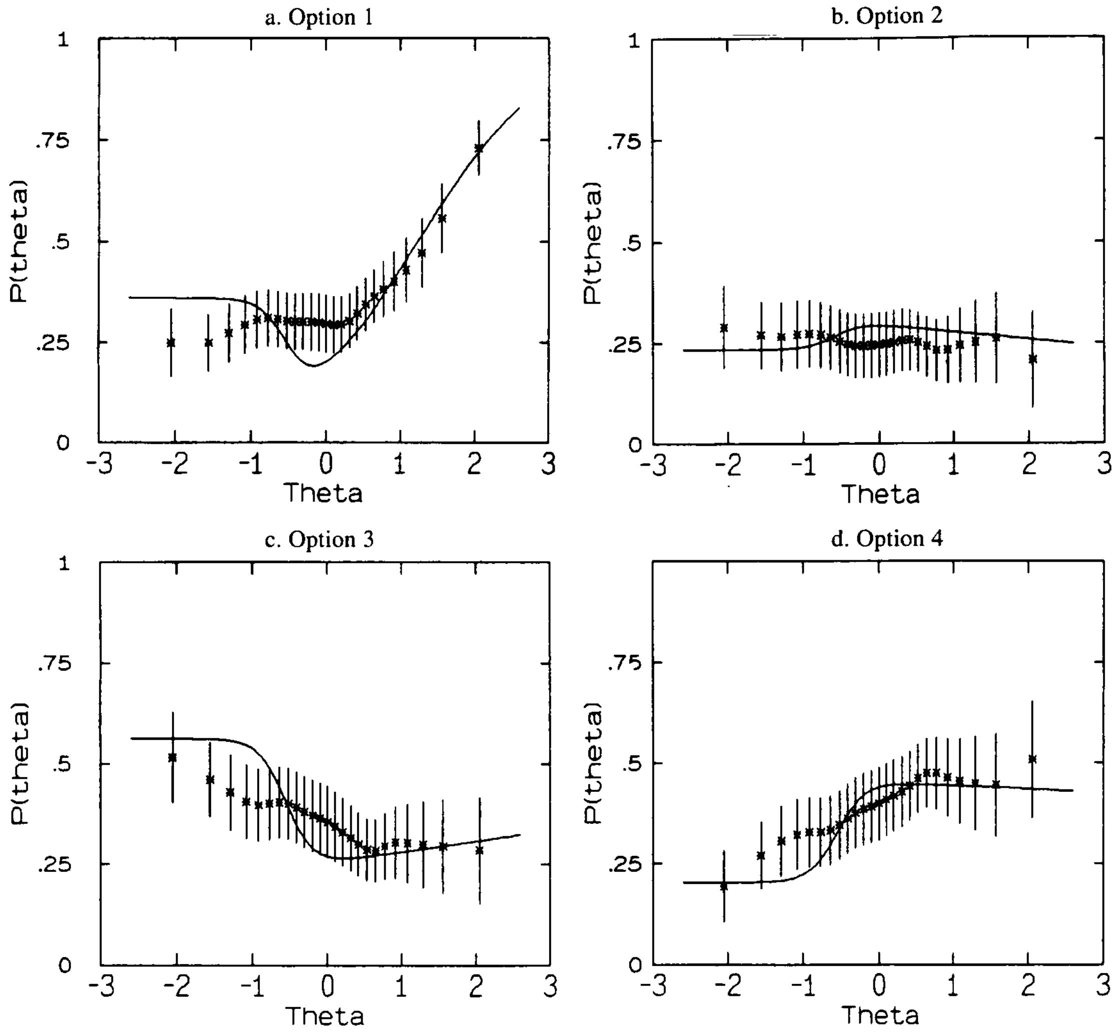
d. Option 4

and divided by the expected frequency, led to a contribution of 7.78 to the item's χ^2 of 12.34.

Oscillations of the sort seen in Figure 2 (i.e., local "hills" and "valleys")—and sometimes considerably larger oscillations—occurred with unfortunate frequency for the TSMCM. Sometimes all of the items in a MULTILOG run had very large oscillations (as noted previously, items were analyzed in batches of approximately 20). Mixing the items in a batch that had many oscillations with items that exhibited fewer oscillations and reanalyzing sometimes reduced or eliminated the problem.

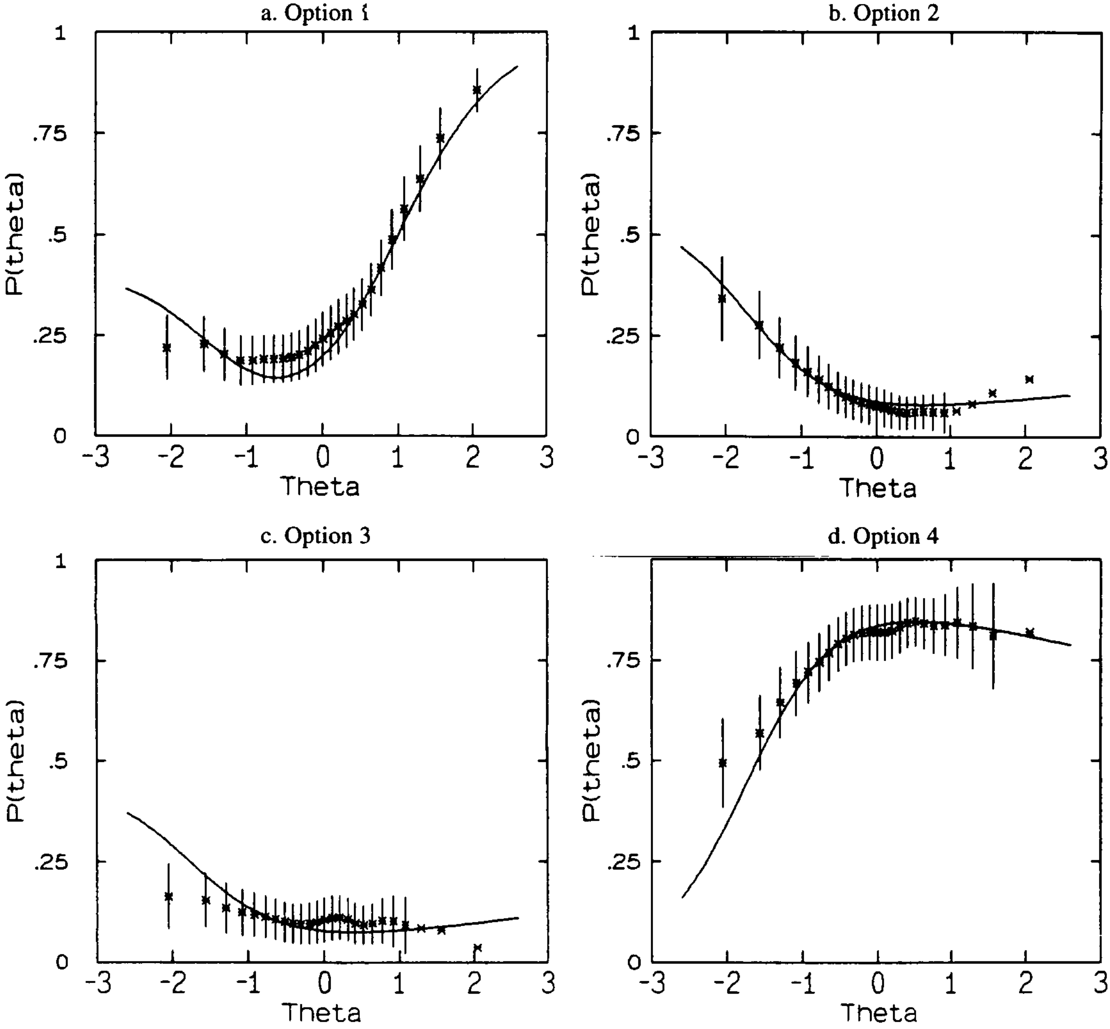
Figures 4 and 5 present CORFs estimated by the MFSM for the same items presented in Figures 2 and 3, respectively. The χ^2s for the MFSM were 4.46 for Item 52 and 2.28 for Item 53. These figures show local hills and valleys, which may be artifacts of the nonparametric estimation method or may indicate multidimensionality in the item pool.

Figure 2
ASVAB Quantitative Item 52 Analyzed by the TSMCM



As noted previously, the χ^2s were generally lower for the ASVAB verbal test than for the quantitative test. A comparison of the fit plots for the quantitative test to the fit plots for the verbal test suggested an explanation: The verbal test was substantially easier than the quantitative test, and it was less difficult to estimate response functions for easy items. Item 11, as analyzed by the SMCM, was typical of the verbal test (see Figure 6). The CORF for the correct option (Figure 6a) was estimated quite accurately. The CORFs for the incorrect options (Figures 6b–6d) were estimated reasonably well in the lower θ range in which examinees had a nontrivial chance of answering incorrectly. In higher θ ranges, virtually no examinee gave an incorrect response. The implication is that the value of the CORF is irrelevant because the ORF is $[1 - P(\theta)]$CORF, which is nearly 0. Consequently, the important part of a CORF was reduced to a nearly straight line for a small range of θ (basically for $-2 < \theta < -1$). Nonlinearities of the CORFs of the sort seen for Options 2 and 4 in Figure 3b and Figure 3d, which were difficult to estimate, were not encountered on easy items.

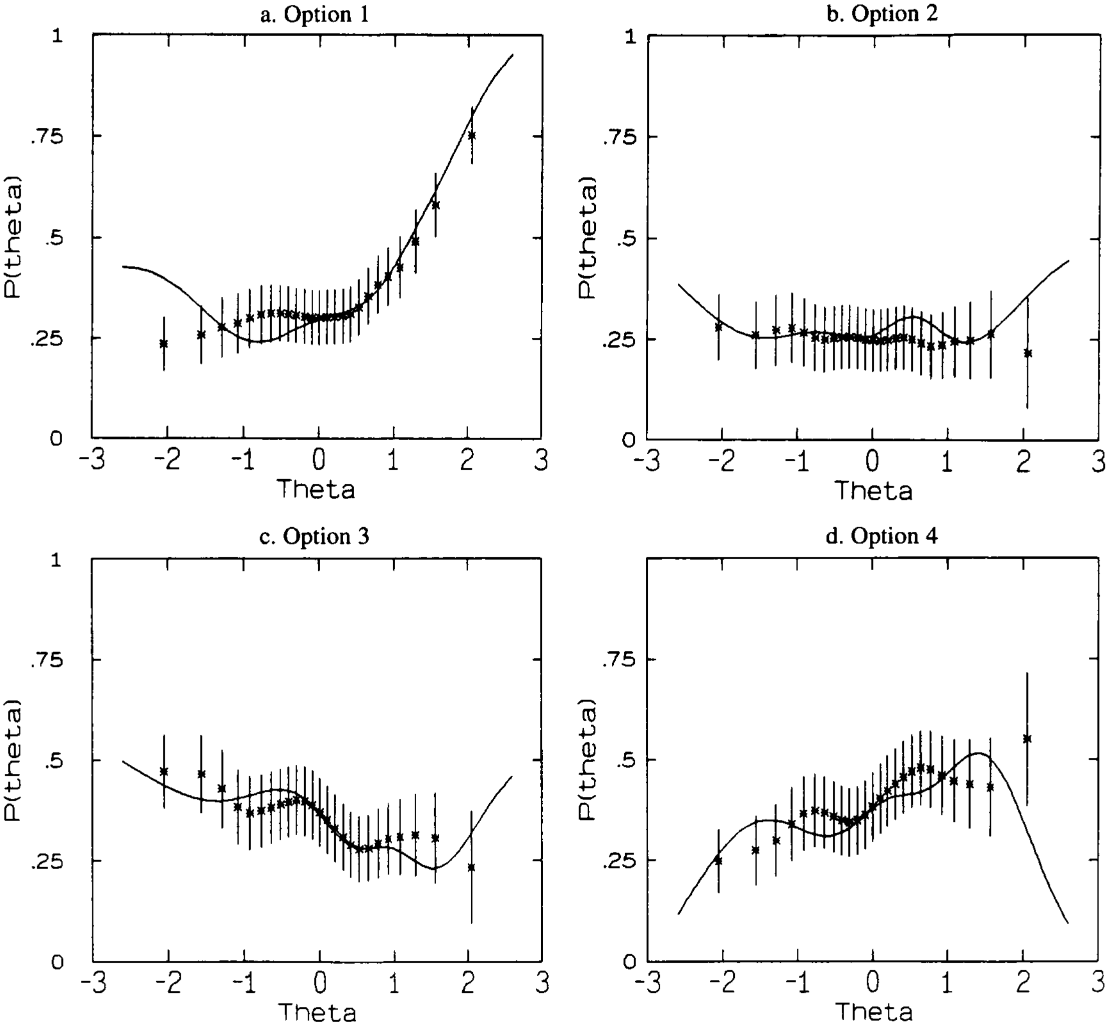
Figure 3
ASVAB Quantitative Item 53 Analyzed by the TSMCM



SAT

Substantial difficulties were encountered when the SAT tests were analyzed with MULTILOG. As noted earlier, MULTILOG did not converge properly when omits were excluded from the likelihood function. When omits were included as a response category, MULTILOG would sometimes "lose θ" and treat omit as the correct response. More specifically, for each item included in a run where this occurred, the omit category would have the largest estimated item discrimination parameter and so, according to the estimated model, the most probable response by high θ examinees would be nonresponse (i.e., an omit). However, the χ^2 fit statistics were very poor for such solutions. Consequently, when θ was "lost" for a particular item set, items were reshuffled into different sets of approximately 20 items and reanalyzed. Reshuffling and reanalyzing continued until solutions in which the correct option consistently had the

Figure 4
ASVAB Quantitative Item 52 Analyzed by the MFSM

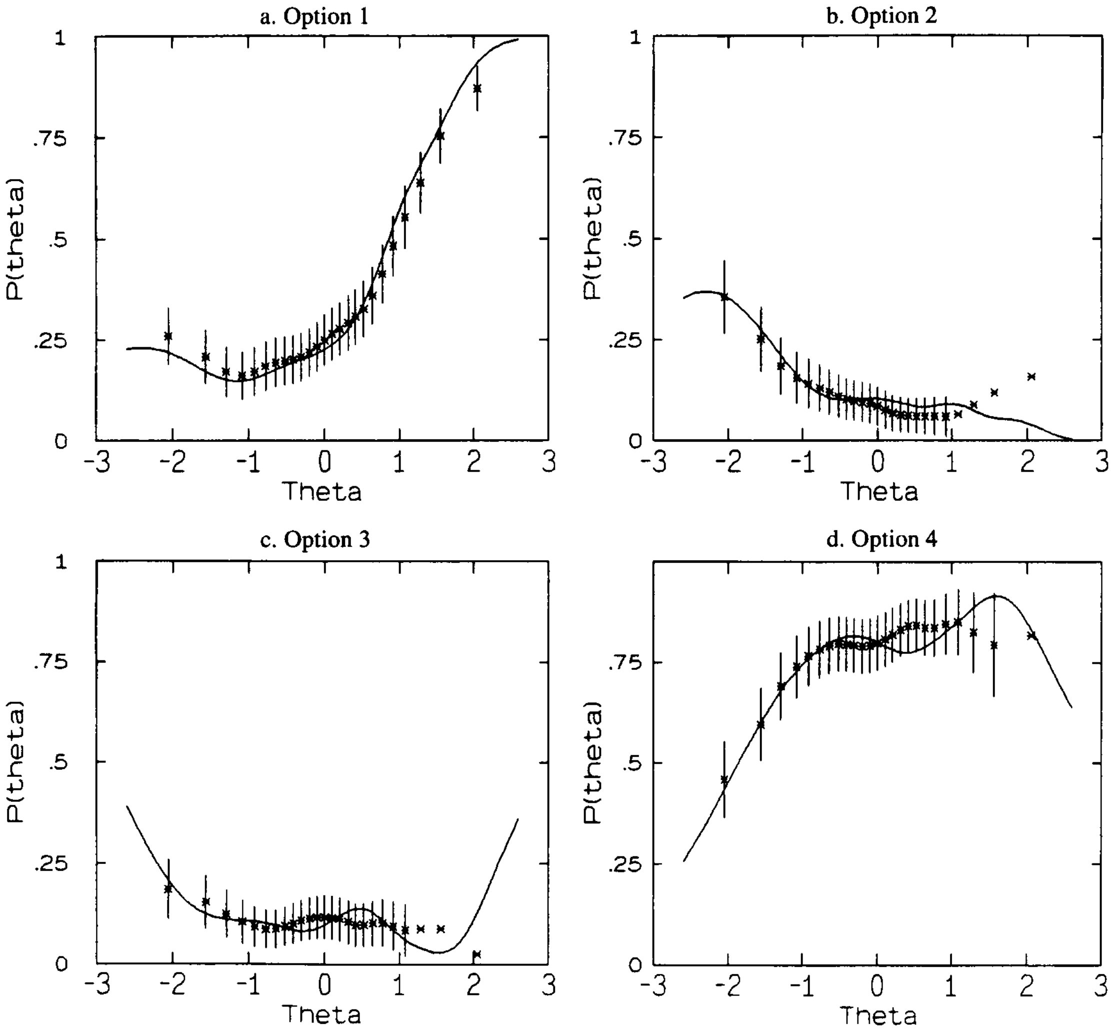

a. Option 1
b. Option 2
c. Option 3
d. Option 4

largest estimated discrimination parameter were obtained. Because each MULTILOG analysis of a set of 20 items took approximately a day (using a 20MHz 386 personal computer with a math coprocessor to perform a maximum of 100 iterations), one to two weeks were needed to analyze each of the SAT tests by either the SMCM or the TSMCM.

For the MFSM analyses of the SAT tests, all 60 SATM items and all 85 SATV items were analyzed in a single run, which took approximately one day on a Hewlett-Packard Series 9000 Model 835 work station that runs jobs in a multitasking environment approximately as fast as a 20MHz 386 personal computer. No convergence problems were encountered.

Table 2 presents the χ^2 statistics for the SAT items. The χ^2s of the SAT items for the BNM were similar to the χ^2s for the ASVAB items presented in Table 1. The fit plots were also very similar: Many items showed small or modest misfit and just a few items showed substantial misfit.

Figure 5
ASVAB Quantitative Item 53 Analyzed by the MFSM



a. Option 1

b. Option 2

c. Option 3

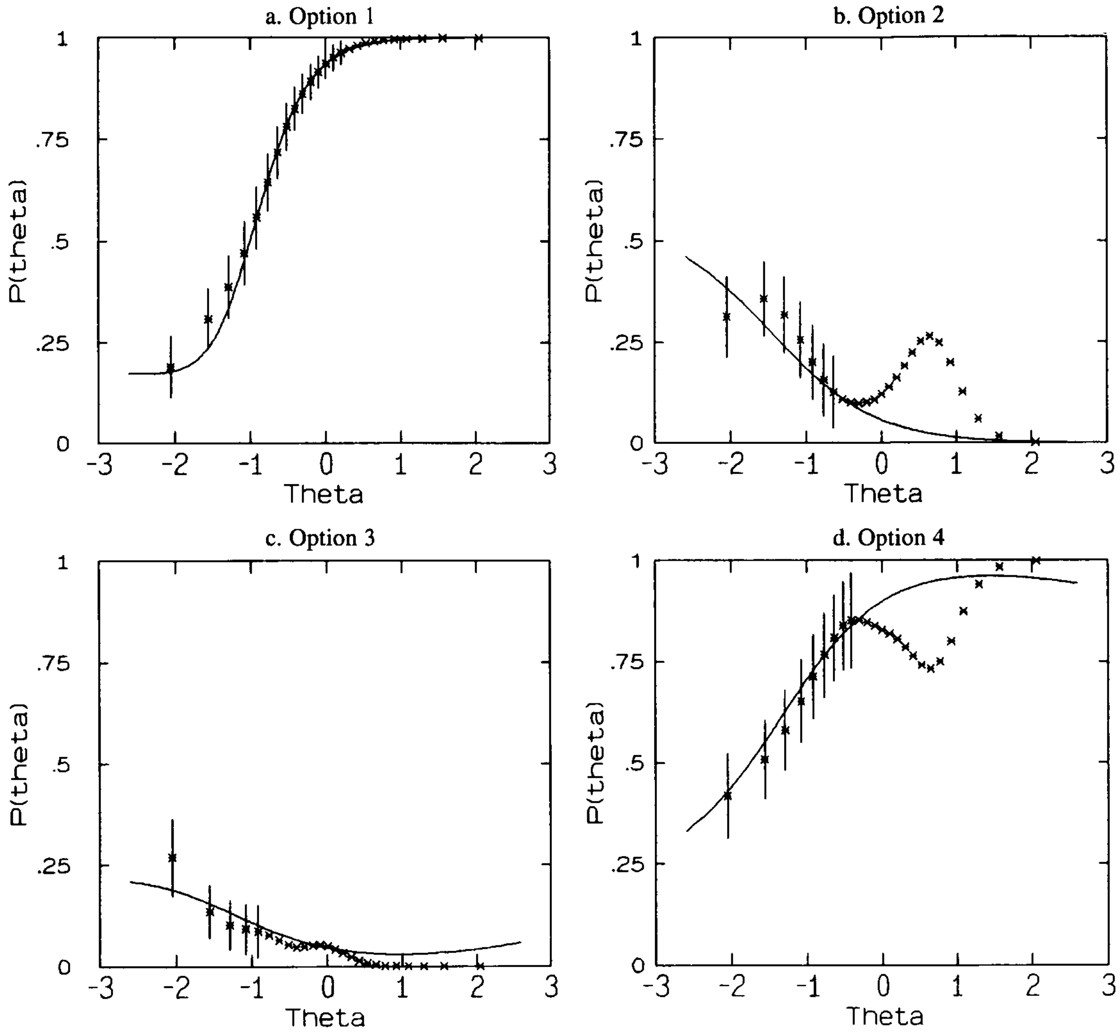
d. Option 4

Table 2 shows that unsatisfactory solutions were obtained for the SMCM and TSMCM despite repeated reanalyses. Figure 7, which contains results for SATV Item 65 analyzed by the TSMCM, shows a typical fit plot.

The χ^2 statistics indicated a fair—but not good—fit for the MFSM when omits were included as a response option. From an examination of response patterns, χ^2s, and the fit plots, it appeared that the unsatisfactory fit for the SMCM and TSMCM, as well as the indifferent fit of the MFSM, were due to a substantial violation of the assumption of local independence. To evaluate this hypothesis, a reanalysis of the SAT tests was performed with the MFSM in which omits were excluded from the likelihood function. Table 2 shows that a much better fit was obtained for the SATV. For the SATM, better fit was obtained for pairs and triples of items.

To further investigate the difficulties caused by omitting on the SAT, an analysis was conducted in which

Figure 6
ASVAB Verbal Item 11 Analyzed by the SMCM


a. Option 1

b. Option 2

c. Option 3

d. Option 4

examinees were grouped according to the number of items they omitted. Relative frequency distributions were computed for the proportion correct [i.e., P_C = (number correct)/(number correct + number incorrect)] for each group. Figure 8 shows that these relative frequency distributions were very similar across groups of examinees who omitted modest to substantial numbers of items. This suggests that the tendency to omit is relatively independent of θ level and a decision to omit may not contribute to the measurement of θ level in the same way as other option choices.

The measure of θ used here—proportion correct on the answered items—might be considered to be an unsatisfactory measure of θ because examinees who frequently omitted might omit difficult items and answer easy items. To evaluate this hypothesis, the proportion-correct statistic was correlated with the three-parameter logistic Bayes modal θ estimate computed from the answered items. The correlation was .97, which indicates that Figure 8 would be relatively unchanged if the proportion-correct statistic were

Table 2
Frequencies, Means, and Standard Deviations (SDs) of SAT χ^2/df Ratios

Test, Model, and Items	Frequency of χ^2/df <1	1–<2	2–<3	≥3	Mean	SD
SATM						
BNM						
Singles	14	23	13	10	1.99	1.27
Doubles	3	13	28	16	2.62	1.27
Triples	0	9	8	3	2.31	.68
SMCM						
Singles	11	14	15	20	3.31	4.31
Doubles	1	7	22	30	3.37	1.68
Triples	0	5	8	7	3.01	1.21
TSMCM						
Singles	3	6	7	44	8.89	10.42
Doubles	0	4	6	50	5.58	3.60
Triples	0	1	3	16	4.12	1.78
MFSM, Omit ORF Estimated						
Singles	11	17	19	13	2.22	1.39
Doubles	2	19	22	17	2.56	1.40
Triples	0	11	8	1	2.19	.83
MFSM, Omit ORF Excluded						
Singles	13	15	15	17	2.37	1.52
Doubles	5	35	16	4	1.89	.75
Triples	0	17	2	1	1.69	.52
SATV						
BNM						
Singles	14	31	17	23	2.25	1.42
Doubles	0	20	36	27	3.04	2.39
Triples	0	12	11	4	2.41	1.11
SMCM						
Singles	7	16	14	48	14.42	33.46
Doubles	0	2	6	75	10.19	11.11
Triples	0	0	1	26	8.44	6.40
TSMCM						
Singles	1	10	10	64	15.24	29.44
Doubles	0	3	10	70	9.37	9.40
Triples	0	1	3	23	6.95	4.64
MFSM, Omit ORF Estimated						
Singles	9	32	24	20	2.91	2.79
Doubles	0	22	23	38	3.63	3.76
Triples	0	7	11	9	3.37	2.81
MFSM, Omit ORF Excluded						
Singles	26	28	14	17	1.89	1.48
Doubles	18	54	10	1	1.44	.54
Triples	7	20	0	0	1.26	.36

replaced with a theoretically preferable IRT θ level estimate.

ACT

Table 3 presents the χ^2 statistics for the ACT Math Usage test. The most striking feature of Table 3 is the very good fit of all four models, which was confirmed by the fit plots. Thus, the results for this test were

Figure 7
SAT Verbal Item 65 Analyzed by the TSMCM

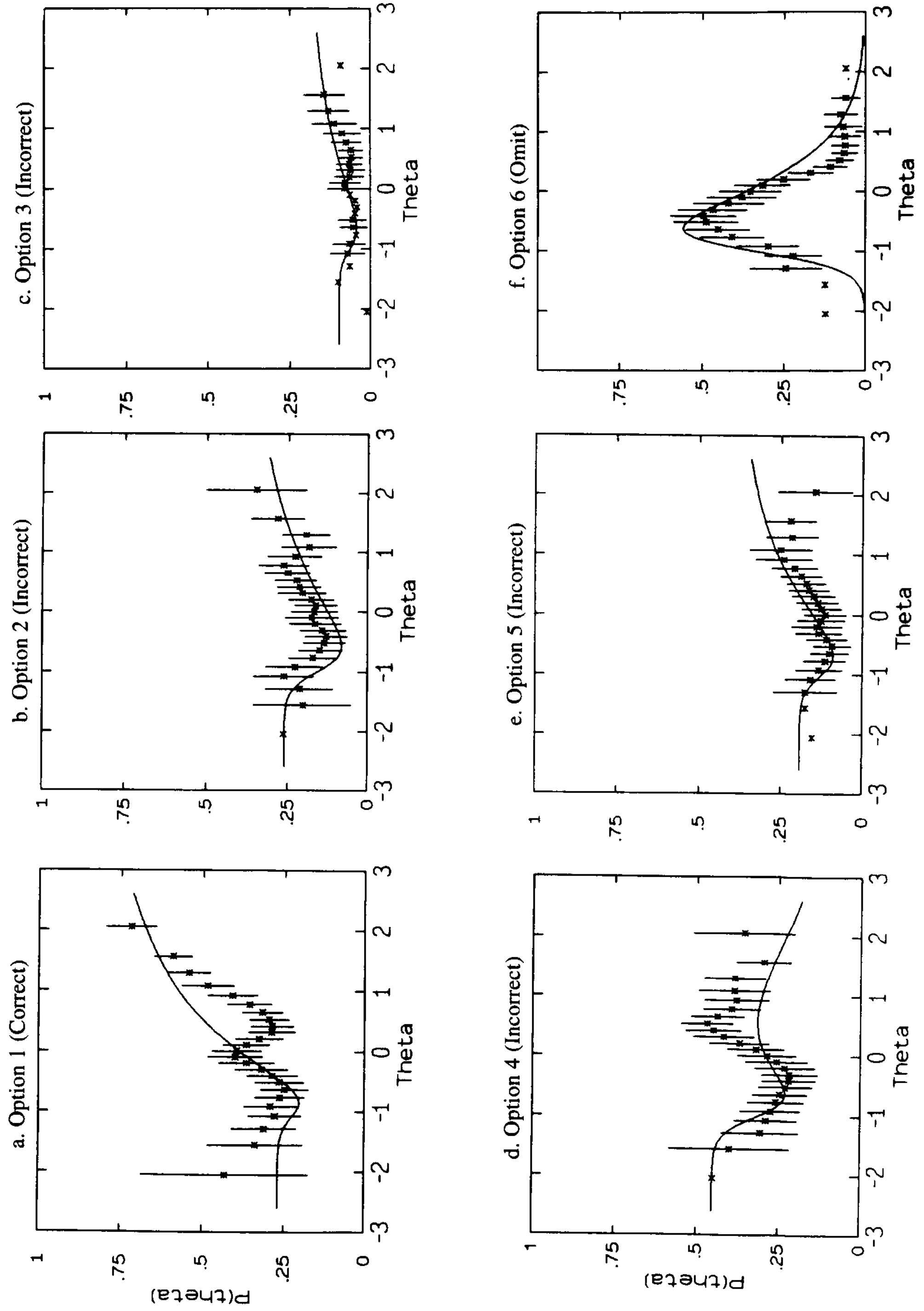
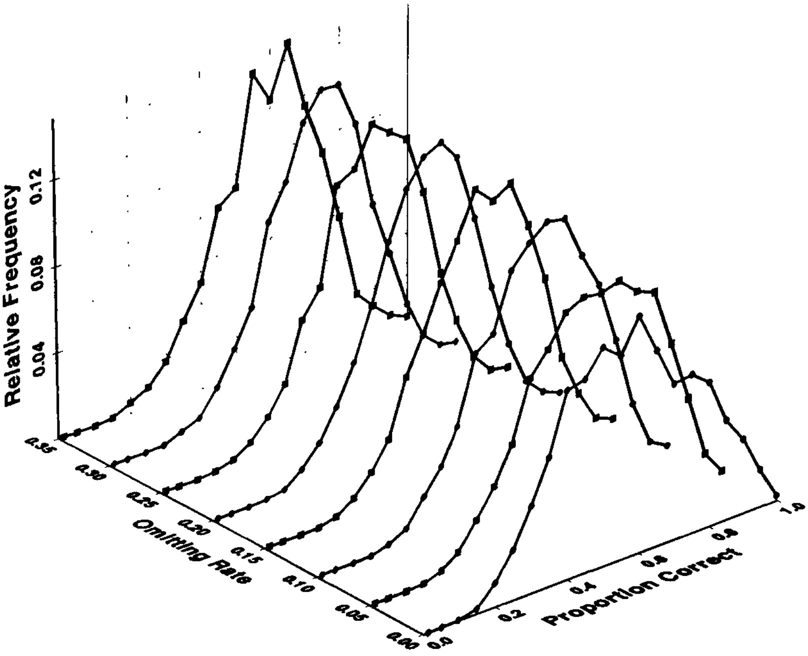


Figure 8
Proportion-Correct Relative Frequency Distributions for Examinees With Different Rates of Omitting



more similar to the results for the ASVAB than for the SAT.

The fit plots revealed some interesting characteristics of the estimation methods. The BNM fit many items fairly well, but had noticeable misfits in the left tails of the CORF for correct options and had substantial difficulty in modeling curvilinearities in CORFs for incorrect options. The SMCM did better than the BNM in modeling the lower asymptotes of correct options, but estimated IRFs were noticeably below the empirical pseudo-proportions for approximately half of the items. Curvilinearities in ORFs also were modeled better by the SMCM, although sometimes the estimated functions were too flat and sometimes the estimated functions showed significant nonlinearities that did not appear in the empirical pseudo-proportions. A different type of problem in modeling lower asymptotes of correct options was apparent for the TSMCM: On approximately one-sixth of the items, the estimated function began to rise steeply at low θs (a "Sympson effect;" Sympson, 1993), so that very low θ examinees were estimated to be two to three times more likely to respond correctly than low to moderate θ examinees. ORFs estimated by the TSMCM were rarely too flat; when there were large estimation errors, the estimated functions were almost always too nonlinear. Finally, the χ^2s for the MFSM were very similar to χ^2s for the SMCM and TSMCM. In addition, the fit plots for the MFSM had some resemblance to fit plots for the TSMCM: There were upward turns in the left tails of IRFs for a small proportion of items, and large estimation errors (when they occurred) were in the direction of too many nonlinearities.

Discussion

The main conclusion from this study is that fitting polytomous IRT models to multiple-choice item responses is more complex than fitting the three-parameter logistic model to dichotomously scored responses. Apparently, polytomous models and their estimation algorithms are sensitive to a kind of viola-

Table 3
Frequencies, Means, and Standard Deviations (SDs) of ACT Math Usage χ^2/df Ratios

Model,	Frequency of χ^2/df					
and Items	<1	1–<2	2–<3	≥3	Mean	SD
BNM						
Singles	10	18	4	8	1.86	1.33
Doubles	5	25	8	0	1.53	.45
Triples	1	11	0	0	1.42	.26
SMCM						
Singles	12	16	7	5	1.67	1.21
Doubles	5	31	2	0	1.41	.40
Triples	1	11	0	0	1.30	.23
TSMCM						
Singles	11	14	9	6	1.78	1.21
Doubles	4	32	2	0	1.43	.34
Triples	1	11	0	0	1.29	.20
MFSM						
Singles	13	13	9	5	1.61	1.13
Doubles	5	31	2	0	1.37	.38
Triples	1	11	0	0	1.25	.22

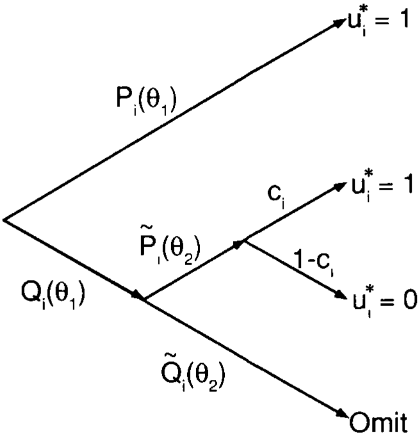
tion of local independence (high omitting rates) that has little effect on dichotomous models. In addition, there is a wider variety of shapes that must be modeled, not just the logistic functions estimated by dichotomous models.

Research by Reckase (1979), Drasgow & Parsons (1983), Harrison (1986), and others has found evidence of considerable robustness of dichotomous model estimation methods to multidimensionality. For example, when there is a general factor underlying responses to all items and several specific factors, each of which affects a subset of items, the θ recovered by LOGIST (Wingersky, Barton, & Lord, 1982) is strongly related to the general factor and almost unrelated to the specific factors. This is ideal if it can be assumed, for example, that on a test of quantitative reasoning the general factor corresponds to mathematical ability and the specific factors correspond to the various content areas assessed by the test (e.g., arithmetic reasoning, algebraic reasoning, geometric reasoning, and so forth). Humphreys (1970, 1981) and Roznowski (1987) made compelling arguments that tests should be constructed according to this general factor/several specific factors paradigm (see Schmid & Leiman, 1957, for a psychometric model for this conceptualization).

The SAT penalizes examinees for incorrect responses. An examination of item response patterns reveals very large individual differences in propensity to omit: Many examinees answer all items and many examinees omit a substantial proportion of items. Because the distributions of the statistic P_C were found to be nearly invariant across different numbers of items omitted (except at the extremes), it appears that omitting propensity is surprisingly independent of θ level. Thus, each SAT test is (at least) two dimensional, with the test measuring omitting propensity as well as an intellective trait. However, this multidimensionality seems fundamentally different from the Schmid & Leiman (1957) conceptualization of multidimensionality to which dichotomous IRT models have considerable robustness.

Figure 9 presents a simple process model for SAT items. Here θ_1 is the trait ordinarily considered to be measured by an SAT test (i.e., verbal or quantitative), and θ_2 is an examinee's propensity to omit. Because the frequency distributions of P_C were very similar for examinees who omitted different numbers of items, it seems reasonable to assume that θ_1 and θ_2 are nearly uncorrelated. In Figure 9, $P_i(\cdot)$ refers to an ordinary two-parameter IRF (based on item difficulty and discrimination) and $Q_i(\cdot) = 1 - P_i(\cdot)$. $\tilde{Q}_i$, however, is an omitting propensity function, which gives the probability of omitting as a function of θ_2 and $\tilde{P}_i(\cdot) = 1 -$

Figure 9
Decision Tree for the SAT Response Process



$\tilde{Q}_i(\cdot)$. Thus, according to Figure 9, an examinee knows the answer to item *i* with probability $P_i(\theta_1)$ and answers correctly. With probability $Q_i(\theta_1)$ the examinee does not know the answer; in this case the examinee decides either to omit the item with probability $\tilde{Q}_i(\theta_2)$, or to answer with probability $\tilde{P}_i(\theta_2)$. Finally, if the examinee decides to answer, a correct response is given with probability c_i and an incorrect response is given with probability $1 - c_i$, where c_i is the probability of guessing correctly given that the examinee does not know the answer to the item. Response functions for incorrect options could be introduced into this model at the final step for a polytomous model.

A cautionary note concerning the fit plots is in order. A fit plot consistently estimates a response function only if the estimated response functions used to compute the posterior densities are correct. In fact, a model with a poor χ^2 can have excellent fit plots. Fit plots are primarily useful for discovering systematic misfit of a few aberrant response functions or a set of items over a particular θ range. This limitation also pertains to more usual fit plots that are computed as regressions on estimated θ level.

The purpose of this research was not to demonstrate which model is best. If the method of estimation and other important aspects of the estimation algorithms are held constant (here marginal maximum likelihood was used for all four models), then the most precise estimates of response functions should be provided by the model that makes the strongest assumptions that accurately characterize the data.

There are at least two ways of making assumptions about the shapes of response functions. The parametric models explicitly use a mathematical form for the response functions, which thereby specifies a class of shapes that the response functions may assume. Alternatively, FORSCORE nonparametrically estimates response functions subject to several constraints. The constraints eliminate unlikely response functions from consideration during estimation, leaving only plausible shapes.

The BNM makes stronger assumptions than either of the other two parametric models, but some of its properties are known to be violated in real data. The SMCM and TSMCM specify response functions in ways that should more closely approximate reality. Nonetheless, according to the χ^2 statistics, neither of the two more complex models fit better than the BNM. One explanation for this finding is suggested by Lord's (1980) paper "Small *N* Justifies Rasch Methods." Unfortunately, the sizes of the calibration samples used here were approximately 3,000; thus, "small *N*" in the context of polytomous measurement may be very large indeed. Alternatively, the disappointing results for the SMCM and TSMCM may be due to sensitivity to violations of local independence. Yet another explanation of these results lies in the fact that several separate

MULTILOG runs were required to calibrate a test due to limitations inherent in the MULTILOG software for the DOS operating system.

A reviewer questioned the relevance of these large sample results for other researchers who typically have much smaller datasets. The focus of this research was not on minimal sample sizes and test lengths needed for polytomous IRT. Instead, the focus was on the range of response function shapes that are needed to model incorrect responses. Because the shapes of ORFs needed to model the tests studied here seem likely to be similar to the shapes needed for other multiple-choice tests, this research has important implications for the selection of a polytomous model. Further research, using simulation methodology, is needed to determine minimum sample size and test length requirements and robustness to violations of assumptions. The study by Reise & Yu (1990) constitutes an important initial contribution in this area.

Of the more complex models, only the MFSM seemed to consistently provide adequate fit. To relate this finding to the estimation algorithms, note that all of the models examined here search a parameter space for a vector of parameters that maximizes the fit to data. All of the models attempt to keep the search space small to improve the efficiency of estimation. Three of the models (the BNM, SMCM, and TSMCM) attempt to keep the search space small by translating psychological intuitions into functional forms with small numbers of parameters. The MFSM attempts a different strategy: Constraints are used to exclude parameter vectors corresponding to functions that have more curvature or variation than can reasonably be expected of ability test items. A systematic study of the effects of the constraints implemented in FORSCORE seems likely to improve the fit of estimated response functions and constitutes a fruitful avenue for further research.

This study provided several clear results and raised even more questions. Among the findings are the surprising robustness of the BNM and generally good fit provided by FORSCORE for the MFSM. Questions remain concerning why the SMCM and the TSMCM did not provide better fit than the BNM, how to model the multidimensional response process underlying the SAT, and whether alternative constraint specifications could further improve the fit of the MFSM response functions.

References

Abrahamowicz, M., & Ramsay, J. O. (1992). Multi-category spline model for item response theory. *Psychometrika, 57,* 5–27.

American College Testing Program. (1988). *ACT assessment program technical manual.* Iowa City IA: Author.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561–573.

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

Department of Defense. (1984). *Armed Services Vocational Aptitude Battery (ASVAB) test manual* (DoD 1304.12AA). Chicago: Military Entrance Processing Command.

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and achievement tests.* New York: College Entrance Examination Board.

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

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Multi-test extensions of practical and optimal appropriateness indices. *Applied Psychological Measurement, 15,* 171–191.

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

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

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91–115.

Humphreys, L. G. (1970). A skeptical look at the factor pure test. In C. E. Lunneborg (Ed.), *Current problems and techniques in multivariate psychology: Proceedings of a conference honoring Professor Paul Horst* (pp. 23–32). Seattle: University of Washington.

Humphreys, L. G. (1981). The primary mental ability. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), *Intelligence and learning* (pp. 87–102). New York: Plenum.

Levine, M. V. (1988, May). *Annual report of progress.* Iowa City IA: Office of Naval Research Contractor's Conference.

Levine, M. V. (1989, August). *Latent trait theory as fundamental measurement.* Paper presented at the meeting of the Society for Mathematical Psychology, Irvine CA.

Levine, M. V. (1993). *Orthogonal functions and the finiteness of continuous item response theories.* Unpublished manuscript.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675–685.

Levine, M. V., & Williams, B. (1991, May). *An overview and evaluation of nonparametric IRF estimation strategies.* Paper presented at the Office of Naval Research Contractors' Meeting on Model-Based Measurement, Princeton NJ.

Levine, M. V., & Williams, B. (1993). *Nonparametric models for polychotomously scored item responses: Analysis and integration.* Unpublished manuscript.

Lord, F. M. (1980). Small *N* justifies Rasch methods. In D. J. Weiss (Ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference* (pp. 386-395). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

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

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG 3.* Mooresville IN: Scientific Software.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133–144.

Roznowski, M. (1987). Use of tests manifesting sex differences as measures of intelligence: Implications for measurement bias. *Journal of Applied Psychology, 72,* 480–483.

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

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

Samejima, F. (1983). Some methods and approaches of estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord* (pp. 159–182). Hillsdale NJ: Erlbaum.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61.

Sympson, J. B. (1986, August). *Extracting information from wrong answers in computerized adaptive testing.* Paper presented at the meeting of the American Psychological Association, Washington DC.

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

Sympson, J. B. (1993). *Extracting information from wrong answers in computerized adaptive testing* (Tech. Rep. No. TN-94-1). San Diego CA: Navy Personnel Research and Development Center.

Thissen, D. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement, 13,* 201–214.

Thissen, D. (1986). *MULTILOG user's guide* (Version 5). Mooresville IN: Scientific Software.

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

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

Williams, B. (1986, April). *The shapes of item response functions.* Paper presented at the Office of Naval Research Model-Based Measurement Contractors Conference, Gatlinburg TN.

Williams, B., & Levine, M. V. (1993). *FORSCORE: A computer program for nonparametric item response theory.* Unpublished manuscript.

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

Acknowledgments

This research was supported primarily by Contract No. N00014-89-K-0059 from the Office of Naval Technology. Additional support was provided by Office of Naval Research Contract No. N00014-86K-0482, NR 442-1546.

Author's Address

Send requests for reprints or further information to Fritz Drasgow, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820, U.S.A. Internet: fdrasgow@uiuc.edu.

A Supplement to "The number of Guttman errors as a simple and powerful person-fit statistic" (Volume 18, Number 4, pp. 311–314)

Rob R. Meijer, University of Twente

Meijer (1994) used a statistic to determine person-fit that was defined as follows:

$$G = \sum_{g=1}^{k-1} \sum_{h=g+1}^{k} f_{gh} , \tag{1}$$

where

g and h are item indexes,

k is the number of items in the test,

f_{gh} = 1 if a person has a 1 (correct, keyed response) on the easier item and a 0 (incorrect, not keyed response) on the more difficult item, and

f_{gh} = 0 otherwise.

G is based on the number of errors from the deterministic Guttman (1950) model. This may have given the impression that this statistic was Guttman's own error definition for his deterministic model. This is not the case. G is the number of errors from the deterministic Guttman model as defined by Loevinger (1947, 1948).

Guttman (1950) defined the number of errors by counting the "... number of responses which would have been predicted wrongly for each person on the basis of his scale score ..." (Guttman, 1950, p. 77), whereas Loevinger (1947, 1948) defined the number of errors by counting all error pairs. A small illustration may clarify the difference.

Assume a test consisting of five items ordered according to increasing item difficulty. Thus, the item score pattern for someone with three 1s according to the perfect Guttman (1950) model is [11100]. A person with the pattern [01011] has four errors according to Guttman's error definition and five errors according to Loevinger's definition. Because Loevinger's definition and scaling approach appeared to be more useful for defining a probabilistic version of the deterministic Guttman model than Guttman's own definition (e.g., Mokken & Lewis, 1982), Loevinger's error definition was used as a simple person-fit statistic.

References

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

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph, 61,* No 4.

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

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417–430.

Acknowledgment

Thanks are due to Bert Green for pointing out this historical inaccuracy.

Author's Address

Send requests for reprints or further information to Rob R. Meijer, TO/OMD, University of Twente, P.O. Box 217, 7500 AE, Enschede, The Netherlands. Internet: Meijer@edte.utwente.nl.