A Generalized Logistic Item Response Model Parameterizing Test Score Inappropriateness

Nancy L. Strandmark and Robert L. Linn
University of Illinois at Urbana-Champaign

The person response curve has been suggested as a possible model for test score inappropriateness (Lumsden, 1977, 1978; Weiss, 1973). The two-parameter person response curve proposed by Lumsden includes a person slope parameter but abandons the notion of differential item relatedness to the underlying trait. As an alternative, a generalized logistic model is considered that includes all item parameters of the three-parameter logistic model (Birnbaum, 1968). In addition to the usual person location parameter, the model has extra person parameters representing two possible characterizations of test score inappropriateness: a slope parameter indicating the degree to which a person responds differently to items of varying difficulty, and an asymptote parameter measuring a person's proclivity to engage in effective guessing or to omit items in the presence of partial information. To assess the model's feasibility, statistical comparisons were made between parameter estimates from data simulated according to the model and the original simulation parameters. The results seem encouraging, but additional empirical study is needed before firm conclusions can be drawn.

An important psychometric problem that has received considerable attention in recent years concerns the accuracy and validity of individual test scores. Traditional indices, such as coefficient alpha and the standard error of measurement, provide descriptive information for the group but completely ignore any variation in measurement error across persons. For the purpose of assessing the appropriateness of individual test scores, methods are needed that quantify measurement error at the level of the individual examinee. It is this problem that is considered here.

Tatsuoka and Linn (1983) have distinguished two general classes of techniques for use in detecting atypical patterns of item responses produced by individuals. The essential difference between the two approaches is that one is based on ordinary descriptive statistics derived from the persons $\times$ items score matrix, whereas the other is model-based. Examples of the first class of measures are the personal biserial correlation (Donlon & Fischer, 1968), Sato's caution index (Tatsuoka, 1979), and the individual and norm consistency indices (Tatsuoka & Tatsuoka, 1982). Model-based indices, all based on item response theory (IRT), include the appropriateness indices of Levine and his colleagues (Drasgow, 1982; Drasgow, Levine, & Williams, 1984; Levine & Drasgow, 1982; Levine & Rubin, 1979), the squared standardized residual (Wright, 1977; Wright & Stone, 1979), and the IRT indices derived by Tatsuoka and Linn (1983) as generalizations of the discrete descriptive statistics of Sato's student-problem curve theory (Tatsuoka, 1979). (Hulin, Drasgow, & Parsons, 1983, have provided an excellent introductory discussion of the rationale for appropriateness measurement, with detailed descriptions of several of the techniques that have been implemented as well as some of the empirical results.)

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 4, December 1987, pp. 355–370

0146-6216/87/040355-16$2.05

Except for the Gaussian model of Levine and Rubin (1979), all of the above IRT-based appropriateness measurement techniques conceptualize test score inappropriateness as the failure of the assumed test model to provide an adequate fit to the item responses of an individual examinee. An alternate strategy is to model the most plausible types of test inappropriateness directly by including additional person parameters within the assumed model. This approach represents an attempt to identify, explain, and individually quantify some of the sources contributing to the lack of model fit, indexed by statistics such as the squared standardized residual (Wright, 1977; Wright & Stone, 1979) and the IRT appropriateness indices (Drasgow, 1982; Drasgow et al., 1984; Levine & Rubin, 1979).

It is this second modeling approach that was taken by Lumsden. Based on ideas suggested by Weiss (1973; Trabin & Weiss, 1979, 1983; Vale & Weiss, 1975), Lumsden proposed the concept of a person response curve as an alternative to the item response curve. According to Lumsden (1977, 1978, 1980), it is much more realistic to localize measurement unreliability within the person rather than within the test item, as is customary. The model he favored, then, is that of a two-parameter person response curve of an examinee's expected item scores regressed onto item difficulty, which is the sole item parameter. Lumsden rejected the assumption of differential item variances implied by a model that includes an item slope parameter because, in his opinion, it precludes the existence of test homogeneity, which is the primary assumption of all IRT models and is necessary to ensure local independence.

Lumsden (1980) and others (e.g., Andrich, 1978) have found it useful to conceptualize a person-item encounter using Thurstone's law of comparative judgment (see Torgerson, 1958). When the variances of both person and item propensity distributions are possibly nonzero and are not necessarily identical, this formulation leads to a model that is similar in spirit to the one introduced in this paper. Both models include parameters for the variances of the two propensity distributions, but they differ with respect to the way in which those parameters mathematically combine. Hence, the Thurstone formulations can be regarded as competitors to the model presented here.

When models such as the Gaussian model, the person response curve, the Thurstone-like model, and the generalized logistic model of this paper are fit to observed data, they automatically produce one or more values containing information about the validity of individual test scores, thus permitting hypotheses to be made regarding plausible sources of invalidity when it is detected. These models use important information that has always been available in test protocols but which has heretofore been ignored. Tapping this information could result in wiser usage of measuring instruments, because only scores of acceptable validity would be used in practical applications of test data.

Generalized Logistic Model

The model proposed here is a generalized item response model that includes both a person parameter and an item parameter; their multiplicative combination determines the slope of the item and person response functions. An additional parameter of this model characterizes differential proclivities of examinees to guess the correct answer when they do not know it. The expected item score of person g on item i is expressed by the equation

$$P_{gi} = (c_i+\gamma_g)+\frac{1-(c_i+\gamma_g)}{1+\exp[Da_i\alpha_g(\theta_g-b_i)]} \quad , \qquad (1)$$

where θ and b are the locations of the person and item on the underlying trait continuum, α and a are person and item parameters that together moderate the slope of the regression function, and γ and c are person and item parameters that adjust the lower asymptote of the function to account for individual guessing behavior on particular test items. The constant D is set equal to 1.702 in order to maximize resemblance of the logistic function to a normal ogive.

Four constraints on the model parameters are needed in order to ensure their uniqueness. Both the origin and unit of the location parameters, θ

and b, must be specified. This can be accomplished by requiring that $\mu_\theta = 0$ and $\sigma_\theta = 1$, which is the same convention that is usually applied for the two- and three-parameter logistic models. Even with these two constraints, the units of the slope parameters, α and a, remain unidentified. After multiplying one of them by σ_θ or by multiplying them by complementary factors of σ_θ, it is still possible to multiply values of α by any positive factor k and multiply values of a by its multiplicative inverse, $1/k$, without affecting model fit.

This fact implies that the units of α and a must be fixed by an additional arbitrary constraint; it is suggested that the sum of squares of the α parameter values be required to equal N. A further lack of uniqueness in the model involves the asymptote parameters, γ and c. If a positive constant, k, in the range (0,1), such that $k \geq \text{Min}(\gamma)$ and $k \leq 1 - \text{Max}(c)$, is subtracted from all values of γ and is added to all values of c, model fit is again unchanged; use of the smallest value satisfying the stated conditions is suggested, which will almost certainly be Min(γ) in practice.

The suggested set of constraints is consistent with the current practice of guaranteeing the uniqueness of model parameters by imposing reasonable but arbitrary conditions on the person parameters. With these four constraints, or alternate ones having equal merit, the problems of the uniqueness and identifiability of parameter estimates are resolved, and the parameters should be estimable, given adequate data.

This model is a generalization of the unidimensional logistic item response models that are in current use. If γ is required to assume a constant value in the range (0,1), it is in effect absorbed into each item asymptote parameter, c. In a similar manner, if α is required to be a negative constant across all persons, it is absorbed into a, with a minus sign attached. With these changes the model reduces to the three-parameter logistic model (Birnbaum, 1968). Requiring either c or γ to be a constant (or constraining both of them at 0) effectively eliminates one or both of those parameters from the model; likewise, constraining a and/or α to positive and negative constant values, respectively, also simplifies the model. Any of the 15 combinations of slope and asymptote parameters can be removed in one of these ways. With the additional possibility of requiring both c and γ to be nonzero constants, there are then a total of 19 possible submodels of Equation 1.

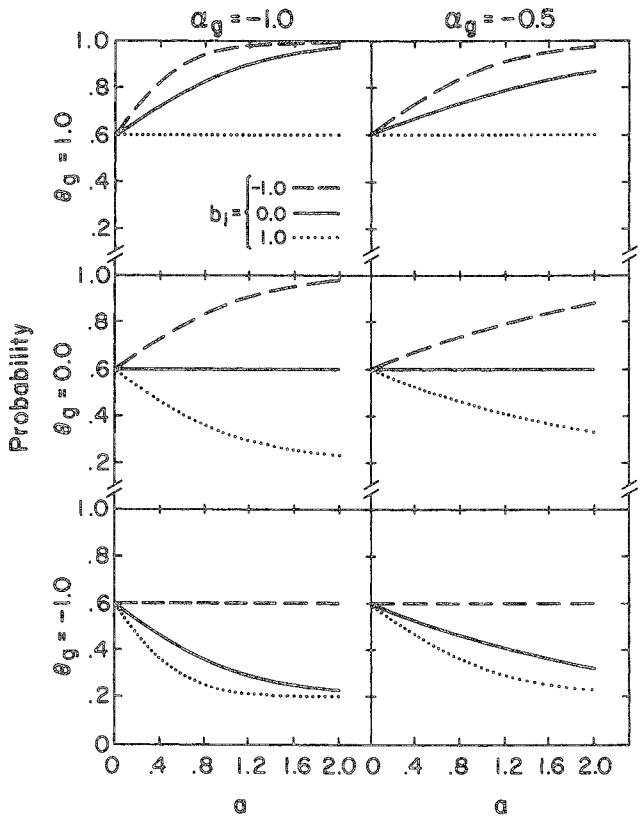
The additional person parameters in this generalized model are an explicit attempt to represent directly two possible characterizations of test score inappropriateness. A unique feature of the model is that it defines a continuous regression surface rather than a simple regression curve. For a fixed person, α, θ, and γ are constant across all items, while a, b, and c are random variables; for a fixed item, the roles of parameters and random variables are reversed. In other words, whether each of these triples is a vector of parameters or of random variables depends on whether it is the person or the item that is considered fixed. When the item is fixed, the equation represents an item response function; when the person is fixed, it describes a person response function.

Plots of Person Response Functions

Whether Equation 1 is momentarily being considered a person or an item response function, it is impossible to depict it in graphic form; doing so would require four dimensions. The best that can be accomplished is to fix all parameters and all but one of the random variables, and then plot probability as a function of values of the remaining random variable. Because examination of such plots can be helpful in clarifying the meaning of this model, a sequence of person response function plots is provided in Figures 1, 2, and 3. Only plots with a and b on the horizontal axis are shown; when probability is regressed onto c, the result is always linear, and corresponding functions differ only in slope and/or separation.

In Figure 1, the probability of a correct response for selected values of the three person parameters is plotted as a function of a. The value of c is also constant, so that the three functions on the same pair of coordinate axes differ from each other only in the value of b, which is either -1, 0, or 1. The

Figure 1
Person Response Functions Regressed Onto a
With a Separate Curve for Each of Three Values of b_i
(c_i = .10 and γ_g = .10)

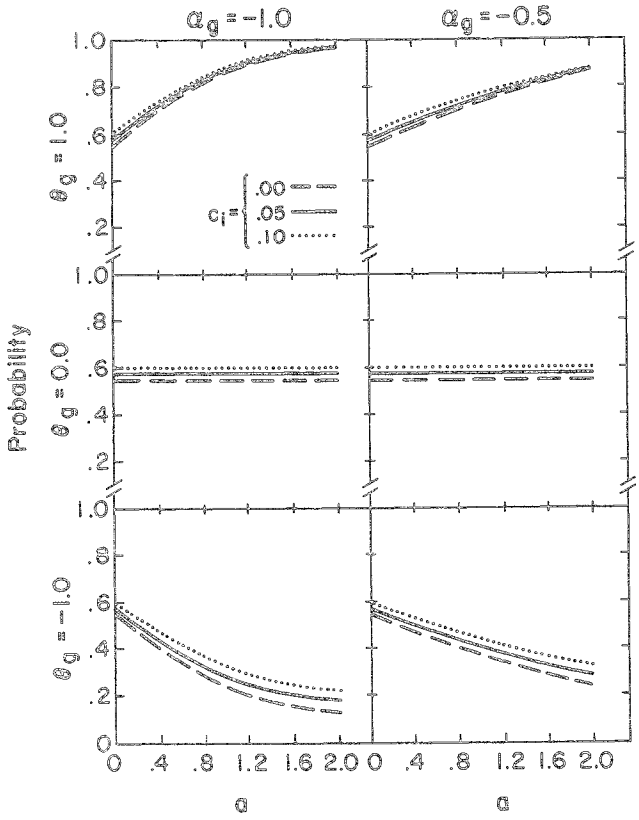


highest curve always corresponds to the lowest value of b, and the lowest curve always corresponds to the highest value. Because b is the item difficulty parameter, this merely reflects the fact that easier items have higher probabilities of correct response than do items of greater difficulty.

In Figure 2, the probability of a correct response for selected values of the three person parameters is plotted as a function of a, as in Figure 1. The value of b is also constant, so that the three functions on the same pair of coordinate axes are alike except for the value of c, which is either .00, .05, or .10. In each plot, the highest curve corresponds to the highest value of c and the lowest curve corresponds to the lowest value, illustrating how the item asymptote parameter reflects the possibility of effective guessing on a test item.

The plots in any one column of Figures 1 and 2 differ from each other only in the value of θ, the person location parameter, which equals 1, 0, and −1 for rows 1, 2, and 3, respectively. Because θ is the usual ability parameter, it is not surprising that increases in its value are accompanied by increases in the value of P_{gi}. The plots in any one row of Figures 1 and 2 differ from each other only in one respect. The person slope parameter, α, is −1.0 in column 1 and −.5 in column 2. Comparing the two columns of Figures 1 and 2 shows

Figure 2
Person Response Functions Regressed Onto a
With a Separate Curve for Each of Three Values of c_i
($b_i = 0.0$ and $\gamma_g = .10$)

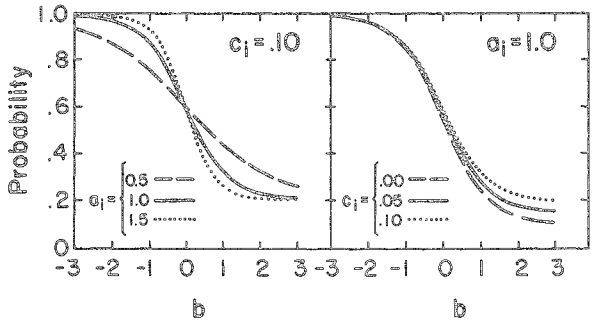


that the higher the value of α, the faster P_{gi} increases or decreases with increases in a, so that α influences the function's rate of change.

Each plot in Figures 1 and 2 illustrates the effect of a, the item slope parameter, represented by the horizontal axis. If $a_i = 0$, α, θ, and b have no influence on P_{gi}; the separation at $a_i = 0$ in each plot of Figure 2 is caused by differences in c. Whenever $a_i = 0$ or $\theta_g - b_i = 0$, then $P_{gi} = (c_i + \gamma_g) + [1 - (c_i + \gamma_g)]/2$. When $\theta_g - b_i > 0$ the curves approach an upper asymptote of 1 as the value of a increases, and when $\theta_g - b_i < 0$ they approach a lower asymptote of $c_i + \gamma_g$ as the value of a increases. In Figure 2, when $\theta_g - b_i > 0$ the amount of separation between curves decreases as the value of a increases, and when $\theta_g - b_i < 0$ the amount of separation increases.

Figure 3 gives two plots of probability as a function of b. Except for negative slopes and asymptotes at the high end of the horizontal axis, the plots do not differ from ordinary three-parameter logistic item response functions. Increasing the value of θ translates a curve to the right, leaving its shape intact, whereas decreasing θ by the same value moves the curve the same distance to the left. From the left panel of Figure 3 it can be seen how increases in a cause the curve to have a steeper slope. The right panel of Figure 3 shows how changes in

Figure 3
Person Response Functions Regressed Onto b
With a Separate Curve for Each of Three Values
of a_i and c_i, Respectively
($\alpha_g = -1.0$, $\theta_g = 0.0$, and $\gamma_g = .10$)



c cause a nonlinear change in the curve, increasing or decreasing probabilities at higher levels of b but having trivial effects at the lower end of the scale. Changes in γ would produce similar results. The slope of each curve in Figure 3 is proportional to the product of a_i and α_g, and the lower asymptote is simply $c_i + \gamma_g$.

As the plots in Figures 1, 2, and 3 illustrate, lower absolute values of α and a lessen an item's success at differentiating among ability levels as well as a person's usefulness in providing information about comparative item difficulties. The asymptote parameters are most influential at low values of θ and at high values of b, where they inflate the probability of correctness. The two parameters combine additively, and therefore do not interact.

Analogous views of item response functions have been omitted but can easily be imagined. Plots with α, the person slope parameter, on the horizontal axis resemble those in Figures 1 and 2. When probability is plotted as a function of θ, the plots are most like those for b in Figure 3, and when γ is on the horizontal axis the plots are linear, just as they are for c. The differences between corresponding item and person response function plots result from the opposite signs of the item and person slope parameters and the reversed relationships of the location parameters to probability (i.e., the greater the value of θ, the more likely it is that the response will be correct, but higher values of b predict lower probabilities of correctness).

The utility of this generalized logistic model depends on the ability of iterative parameter estimation procedures to obtain sufficiently accurate estimates of the person parameters, as well as on the model's ability to capture some of the sources of individual differences contributing to the appropriateness of test scores. An attempt has been made to estimate the parameters of the model using simulated data with what appears to be relatively good success.

Method

Data Generation and Parameter Estimation

Data were generated to simulate the responses of 2,195 persons to 111 test items. The location parameters, θ and b, were both drawn from the standard normal distribution, truncated in the range $(-3,3)$. All other parameters were taken from uniform distributions. The ranges for the slope parameters, α and a, were $(-1.5,-.5)$ and $(.5,1.5)$, respectively, and the lower asymptote parameters, γ and c, were both drawn from the range $(.025,.15)$.

The simulated item score data were submitted to a parameter estimation routine designed to es-

timate all $3n + 3N$ parameters of the model, where n is the number of simulated items and N is the number of simulated examinees. Item scores of only the first 1,000 simulated examinees were included in the calibration sample, so that a total of $3(111 + 1000) = 3{,}333$ parameters were estimated.

An iterative parameter estimation procedure was to be used which, considering the number of parameters, was expected to be comparatively expensive even by IRT standards. For this reason, it was important that the initial estimates provided to the routine be as accurate as possible, assuming that the true parameters were unknown. Initial estimates for the location and slope parameters were computed as functions of the item and person biserial correlations, as outlined by Lord (1980, pp. 33–34). Although Lord's rationale does not strictly apply to this model, it was nevertheless accepted as a reasonable approach. Estimates for γ and c were initially set at .10. An alternate strategy for deriving initial parameter estimates would have been to obtain them by applying LOGIST (Wood, Wingersky, & Lord, 1976) to the persons × items score matrix, and then to its transpose. Neither procedure is exactly suited to this task; therefore the less costly alternative was used.

The person and item parameters were estimated by a cyclical, modified maximum likelihood procedure similar to that used by LOGIST (Lord, 1980; Wood et al., 1976). During the first part of each major cycle, person parameters were reestimated while considering the current item parameter estimates as the true values; after all person parameters were reestimated, the item parameters were updated, with the latest person parameter estimates treated as true parameters. Each person and item was recalibrated by a separate call to the optimization algorithm. This estimation procedure relies heavily on the assumption of local independence of items and persons. It is consistent with current practice for estimating the parameters of two-way item response models for which conditional estimation is not possible. It is not known whether such a procedure converges to a global optimum for a two-way model, but the results generally appear to be good.

The IMSL (1982) subroutine ZXMIN was employed to determine the set of parameter estimates that would optimize the negative log likelihood corresponding to each person or test item. The algorithm is based on the Harwell library routine VA10A (Fletcher, 1972). It uses a quasi-Newton estimation method and is highly regarded as an efficient, accurate, and readily available tool for that purpose (Michael Levine, personal communication, Fall 1982). Even though ZXMIN assumes the existence of the gradient vector and the Hessian matrix, it does not require the user to supply formulas for them; therefore, the only explicit function it requires is the function that is to be optimized.

Before each call to ZXMIN, some of the current parameter estimates that were to undergo modification were transformed by monotonic functions designed to impose constraints on the resulting estimates. The inverse transformations were applied before each evaluation of the negative log likelihood to obviate having to adjust the likelihood function by the Jacobian matrix associated with the particular transformations that were used. (The inverse transformations were applied again after leaving ZXMIN, although, in retrospect, this extra step was probably avoidable.) Imposing bounds on the parameters using such transformations places restrictions on the parameter space, making the estimation procedure only a pseudo-maximum likelihood method. This modification can be conceptualized as a Bayesian-like procedure that is based on informed expectations about the true parameter space. Its primary purpose is to ensure that all estimated parameters fall within reasonable ranges. (This procedure was suggested by Michael Levine, personal communication, Fall 1982.) The function

$$x_1 = \ln\left(\frac{2 - a^*}{a^* - .3}\right) \tag{2}$$

and its inverse,

$$a^* = \frac{1.7}{1 + \exp(x_1)} + .3 \tag{3}$$

were used, with $a^* = a_i$ or $a^* = |\alpha_g|$, depending on whether item or person parameter estimates were

being updated. If $a^* > 2 - \delta$ or $a^* < .3 + \delta$, where $\delta = .0001$, it was replaced by that value before x_1 was computed. Likewise, if the new value of x_1 supplied by ZXMIN was less than ln (.00005) or greater than ln (17,000), it was replaced by that value to prevent underflow and overflow errors. These values were derived by rounding the bracketed part of Equation 2 after it was computed using values of $2 - \delta$ and $.3 + \delta$. For α, the value of a^* from Equation 3 was also multiplied by -1. The slope parameters, a and α, were in this way bounded by (.3001,1.9999) and (−1.9999, −.3001), respectively. In a similar manner, bounds of (.0001,.1499) were imposed on the estimates for c and γ using the equation

$$x_3 = \ln\left(\frac{.15 - c^*}{c^*}\right) \tag{4}$$

and its inverse,

$$c^* = \frac{.15}{1 + \exp(x_3)} \quad . \tag{5}$$

In this case $c^* = c_i$ or $c^* = \gamma_g$ unless $c^* > .15 - \delta$ or $c^* < \delta$, when it was set to the closer of these two values. As before, $\delta = .0001$. If $x_3 < \ln(.0007)$ or $x_3 > \ln(2{,}500)$, the value of x_3 was reset to the value in the inequality. The location parameters were unbounded, so that $x_2 = b^*$, where $b^* = b_i$ or $b^* = \theta_g$, depending on which parameters were being updated.

Due to the complexity of the estimation problem, some compromises were made in choosing specific options for the estimation algorithm. In particular, ZXMIN permits several options for determining initial values of the Hessian matrix. Originally the program specified that ZXMIN compute an estimate of the Hessian matrix. When this procedure failed to enact changes in the parameter estimates, the Hessian matrix estimated by ZXMIN was constrained to be diagonal. Because this procedure provided updates that continually improved from one cycle to the next, it was retained. An alternative that was not tried would have been to estimate the second derivatives independently of the minimization subroutine and then supply the Hessian matrix to ZXMIN. It is not known whether such a procedure would have resulted in better estimates and/or faster convergence.

Parameter estimation required a considerable number of continuation runs. At the end of each run, location and slope estimates were normalized by the following transformations:

$$\hat{\theta}_g^* = \frac{\hat{\theta}_g - \hat{\mu}_\theta}{\hat{\sigma}_\theta} \quad , \tag{6}$$

$$\hat{b}_i^* = \frac{\hat{b}_i - \hat{\mu}_\theta}{\hat{\sigma}_\theta} \quad , \tag{7}$$

$$\hat{\alpha}_g^* = (\hat{\sigma}_\theta)^{1/2}\hat{\alpha}_g \quad , \tag{8}$$

and

$$\hat{a}_i^* = (\hat{\sigma}_\theta)^{1/2}\hat{a}_i \quad . \tag{9}$$

The parameters $\hat{c}_i$ and $\hat{\gamma}_g$ were left unchanged.

Parameter estimation was considered complete when the changes in the person and item likelihoods from one cycle to the next were both less than or equal to .001. A very strict convergence criterion was used for calls to ZXMIN so that all possible values of the original parameters would be estimated with acceptable accuracy, despite the use of the nonlinear transformations that were substituted as input to the estimation subroutine. Thus, it was no surprise that the between-cycles criterion was satisfied first.

The estimation procedure required 68 cycles to reach convergence. Twice during estimation, the parameter estimates began to diverge; whenever this happened, the problem was relieved by backing up to the previously saved estimates and reestimating the updates using a lower value for the maximum number of calls to ZXMIN. A self-contained algorithm designed especially for this particular estimation task might be expected to achieve more accurate estimates with much greater efficiency. The purpose of this study, however, was to examine the feasibility of this model rather than to develop an optimal estimation procedure.

Evaluation of Parameter Estimates

Several different types of descriptive statistics were computed to assess the accuracy of the final parameter estimates by comparing them with the

true parameters from which the data were generated. Before some of the computations were made, additional transformations were carried out on both the simulated and estimated parameters. In particular, the slope parameter estimates were transformed to have equal mean sums of squares and the asymptote parameters were transformed to equalize their means. Different constants from those used for the simulated parameters were required for the estimated parameters. All transformations were consistent with the conditions stated above and were of little consequence; they were used simply because these indeterminacies exist in the model, and it was known that α and a were simulated from distributions with equal dispersions and that γ and c were simulated from distributions with equal means. In the remaining text, all equations will be written without the asterisks; the symbols should then be regarded as either the normalized estimates or the normalized estimates after submission to this additional set of transformations, depending on the context.

Correlations. The simplest of the descriptive statistics are product-moment correlations between the estimated and simulated parameters. Because all valid parameter transformations are linear, they have no effect on the values of these correlation coefficients.

Root mean squared and mean absolute differences. To assess the accuracy of the asymptote parameter estimates, two useful statistics are the root mean squared difference (RMSD) between the original and estimated parameters,

$$\text{RMSD} = \left[\frac{\sum_{i=1}^{n}(x_i-\hat{x}_i)^2}{n}\right]^{1/2}, \tag{10}$$

and the mean absolute difference (MAD),

$$\text{MAD} = \frac{\sum_{i=1}^{n}|x_i-\hat{x}_i|}{n}. \tag{11}$$

Unlike the correlation coefficient, low values of these statistics signify good fit. Also, note that they are not invariant with respect to most valid transformations of the parameters; the only exception is when the constant used to transform the asymptote parameters is the same for simulated and estimated parameters.

Bias statistics. In addition to the criterion that simulated and estimated parameters be highly correlated, it is also desirable that the estimates be unbiased in a statistical sense, so that, on the average, parameters are neither over- nor underestimated. The customary statistic for this purpose is

$$\text{Bias} = \frac{\sum_{i=1}^{n}(x_i-\hat{x}_i)}{n}. \tag{12}$$

Whenever parameter transformations can affect values of the RMSD and MAD statistics, they can also affect the computed bias.

Evaluation of Estimated Probabilities

For purposes of evaluating the information provided by the parameters on the type and extent of individual differences and item properties, the above comparisons are suitable. However, it is also important to examine how successfully the data are fit in the calibration sample, as well as how satisfactorily the model operates across the entire range of probable person and item parameters. To answer such questions, it is necessary to compare the true and estimated probabilities generated by the model. Note that any valid transformations carried out on the simulated and estimated parameters do not affect these statistics.

Root mean squared differences. The first comparison using probabilities involved computation of RMSDs between probabilities derived using the simulated and estimated parameters. Any results from this statistic can be expected to be sample-specific, in the sense that they are generalizable only to a population with similar distributions of the person and item parameters. In addition, each item (person) statistic reflects error not only in the parameters of the item (person) it represents, but also in the estimated person (item) parameters; hence it is not possible to pinpoint the exact causes of inaccuracy. The distributional characteristics of the RMSD are unknown, but it can at least be interpreted by comparison with its perfect baseline of 0 and

can also be used to make ordinal comparisons among items.

Although computing MADs would also have been appropriate, this was not done. Squaring deviations increases the relative contribution of outliers, so that MADs would generally be lower than the RMSD statistics.

Root integrals. Another statistic similar to the RMSD is the root integral

$$\mathrm{RI}_i = \left(\int_{\alpha} \int_{\theta} \int_{\gamma} \Delta P_i^2 \, \partial\gamma \, \partial\theta \, \partial\alpha \right)^{1/2} \quad (13)$$

and its analogue,

$$\mathrm{RI}_g = \left(\int_{a} \int_{b} \int_{c} \Delta P_g^2 \, \partial c \, \partial b \, \partial a \right)^{1/2} , \quad (14)$$

where $\Delta P_i = P_i(\alpha,\theta,\gamma) - \hat{P}_i(\alpha,\theta,\gamma)$ for Equation 13 and $\Delta P_g = P_g(a,b,c) - \hat{P}_g(a,b,c)$ for Equation 14. Unlike the observed RMSDs, these statistics implicitly consider discrepancies in all areas of the range of integration as being of equal significance.

Equations 13 and 14 were calculated using probabilities computed from the simulated and estimated parameters of this study after all parameter transformations had been carried out as described above. The IMSL (1982) subroutine DMLIN, which integrates functions of several random variables by numerical methods, was used. The functions were integrated across the same intervals from which the parameters were originally drawn:

a: (.5,1.5)
α: (−1.5,−.5)
b, θ: (−3,3)
and
c, γ: (.025,.15).

The RMSDs and the root integrals produce different orderings of the persons and items. This is because of the dependence of the RMSDs on the parameter distributions, and the dependence of the root integrals on the ranges of integration. No comparisons between the sets of values of the two statistics can logically be made, therefore, because they provide two distinct kinds of information.

Results

Comparisons Between True and Estimated Parameters

All computed statistics used to assess the parameter estimates are displayed together in Table 1. The bias statistics computed before and after transformation differed enough so that both sets of values are reported.

Correlations. Each of the correlations between estimated and true parameters is statistically significant at $p < .05$. However, the suitability of statistical tests on these correlations is doubtful. Except for the two location parameters, which were sampled from normal distributions and estimated from the simulated data with no bounds imposed, the parameters do not meet the necessary assumptions of normality and homoscedasticity. In addition, the parameters are interdependent, having been estimated together from a common set of data.

Root mean squared and mean absolute differences. The correlation coefficient is a more suitable statistic for the location and slope parameters; therefore, only RMSD and MAD values for the asymptote parameters are given in Table 1. Parameter transformation effected a negligible change in

Table 1
Statistical Comparisons Between True and Estimated Parameters

Statistic	a	b	c	α	θ	γ
Correlation	.781	.967	.178	.552	.873	.166
RMSD	--	--	.039	--	--	.046
MAD	--	--	.033	--	--	.037
Bias						
Normalized	-.223	-.016	-.005	.241	-.028	.002
Transformed	-.196	-.016	-.002	.268	-.028	-.001

only one value, hence only the statistics computed using the original, normalized parameter estimates are reported. Because of the bounds imposed during parameter estimation, statistical tests are inappropriate. However, all of these values can be evaluated with respect to the range of c and γ, which was (.00,.15); by that standard they are relatively good.

Bias statistics. Because the location parameters were sampled from the standard normal distribution and the person location parameter estimates were transformed by Equation 6, the discrepancy of this statistic from 0 for person location parameters reflects only the inexactness in drawing a finite sample from an infinite population. Bias statistics for item location and for person and item slope and asymptote parameters carry more meaning. The size of the bias statistic for the item location parameter estimates is trivial. Person and item slope parameters both tended to be biased outward, and the bias was usually more extreme for persons than it was for items. If upper bounds had not been enforced, this effect might well have been even more drastic. The asymptote parameter estimates, though relatively inexact, seem to be unbiased despite the imposition of bounds.

The statistics cited thus far indicate that item parameters were estimated better than person parameters; but because the distributions from which corresponding parameters were drawn were identical, this can easily be explained by the unequal sample sizes. The location parameter estimates are quite accurate, slope estimates are moderately good, and estimates of asymptotes are only mediocre.

Comparisons Between True and Estimated Probabilities

Root mean squared differences. A RMSD statistic was calculated for each item, averaging across persons, and for each person, averaging across items. For items, the average RMSD was .066, the standard deviation was .011, and the skewness was .961; comparable values for persons were .062, .023, and 1.231, respectively. The five highest and lowest values for items and persons are shown in Tables 2 and 3. It is quite encouraging that the value of the RMSD exceeds .10 for just 2/111 = 1.8% of the items and for only 72/1000 = 7.2% of the persons. However, it must be kept in mind that these values are somewhat reduced by the ideal condition that the person and item location parameter distributions were normal with equal means and variances. In other words, the test was centered exactly where the most information was needed. Among the tabled values, high RMSDs are associated with

Table 2
The 5 Highest and 5 Lowest
Item Root Mean Squared Differences
(n = 111 Items)

Item	a	$\hat{a}$	b	$\hat{b}$	c	$\hat{c}$	RMSD
Highest							
14	.508	1.253	.844	.417	.073	.037	.1134
109	.825	1.270	.209	.018	.049	.100	.1041
68	1.082	1.257	.280	.018	.110	.098	.0986
36	1.432	1.108	1.379	1.155	.032	.077	.0893
96	.806	1.189	.096	.036	.027	.100	.0836
Lowest							
9	.677	.915	-.629	-.276	.054	.094	.0472
80	.616	.775	.100	.167	.085	.081	.0465
35	.527	.752	-.427	-.188	.069	.099	.0440
107	.540	.640	.368	.233	.054	.032	.0438
20	.545	.643	.251	.148	.135	.086	.0405

Table 3
The 5 Highest and 5 Lowest
Person Root Mean Squared Differences
(N = 1,000)

Person	α	$\hat{\alpha}$	θ	$\hat{\theta}$	γ	$\hat{\gamma}$	RMSD
Highest							
111	-.836	-.972	-.965	-.043	.032	.100	.2235
604	-1.460	-1.997	-.525	.041	.110	.080	.1591
251	-1.479	-1.867	-1.050	-.956	.098	.015	.1572
424	-1.405	-.600	-1.224	-1.992	.114	.029	.1484
42	-1.306	-.537	-.030	.122	.084	.094	.1472
Lowest							
260	-.695	-.911	.750	.464	.059	.101	.0265
441	-.768	-1.057	-.043	.020	.064	.069	.0264
391	-.577	-.852	.246	.153	.072	.085	.0262
174	-.518	-.548	.398	.325	.104	.104	.0250
319	-.690	-.820	-.133	-.005	.124	.098	.0228

extreme values of the true slope parameter about half of the time. Also, the true value of b is usually positive and the true value of θ is usually negative for the tabled entities with high values of RMSD.

Root integrals. For items, the average root integral was .064, the standard deviation was .029, and the skewness was .610; corresponding values for persons were .071, .038, and .834, respectively. The five highest and lowest values of these item and person root integrals are shown in Tables 4 and 5. Of the item root integrals, 12/111 = 10.8% are greater than .10; 198/1000 = 19.8% of the person root integrals exceed that value. The root integrals tend to be especially high when both the location and the slope parameters are estimated poorly. When both of those parameters are estimated well, that fact is reflected in a low root integral value.

The item and person root integrals appear to serve the purpose for which they were intended; the high values in Tables 4 and 5 are for items and persons with particularly badly estimated parameters. The fact that the mean item root integral is lower than the mean for persons agrees with expectation based on the greater accuracy of item parameter estimates. This result stands in contrast to that for the two RMSD means.

Discussion

Theoretical Considerations

The ideas incorporated into the generalized logistic model presented here are not entirely new. The person asymptote parameter is meant to recognize known individual differences in the tendency to guess at items that are too difficult, rather than to omit them or to omit items in the presence of partial knowledge. Lord and Novick (1968, p. 304) have used the term "omissiveness" for this particular response style trait. The person slope parameter comes directly from recent thinking involving the concept of a person response curve (Lumsden, 1977, 1978, 1980; Trabin & Weiss, 1979, 1983; Vale & Weiss, 1975; Weiss, 1973).

A useful theoretical comparison involves the contrast between the Gaussian model of Levine and Rubin (1979) and the generalized logistic model of this study. Levine and Rubin were very adamant about their preference always to consider their model as that of an item response function with fixed item parameters and a single random variable representing the individual's trait standing. At each item administration a value of θ_{gi} is sampled from a distribution that is assumed to be normal with mean μ_θ and variance σ^2_θ that are estimable from the data.

Table 4
The 5 Highest and 5 Lowest Item Root Integrals
(n = 111)

Item	a	$\hat{a}$	b	$\hat{b}$	c	$\hat{c}$	RI
Highest							
85	.717	1.198	-1.195	-.466	.087	.146	.1458
14	.508	1.253	.844	.417	.073	.037	.1433
11	1.428	1.710	-1.334	-.895	.033	.100	.1427
56	.760	.770	-2.134	-1.603	.053	.149	.1173
86	.737	1.084	-.989	-.466	.058	.128	.1173
Lowest							
7	1.343	1.141	-2.187	-2.309	.136	.150	.0250
62	.686	.892	1.399	1.135	.097	.150	.0227
66	.905	.956	.254	.238	.037	.079	.0185
46	.614	.669	.275	.109	.077	.092	.0170
5	1.363	1.473	1.884	1.812	.079	.088	.0088

Table 5
The 5 Highest and 5 Lowest Person Root Integrals
(N = 1,000)

Person	α	$\hat{\alpha}$	θ	$\hat{\theta}$	γ	$\hat{\gamma}$	RI
Highest							
507	-1.335	-.771	-1.760	-5.442	.029	.081	.2734
455	-.727	-.352	-1.916	-9.491	.056	.133	.2323
434	-1.189	-1.997	-1.117	-.278	.082	.009	.2289
250	-.742	-1.997	1.287	.443	.063	.099	.2009
886	-.671	-1.370	-1.343	-.466	.067	.029	.1887
Lowest							
775	-.791	-.796	-.210	-.163	.077	.078	.0069
405	-.620	-.602	.560	.469	.111	.103	.0065
497	-.714	-.713	.000	.003	.086	.097	.0065
48	-.631	-.662	.712	.694	.092	.099	.0045
746	-1.419	-1.393	.353	.307	.100	.100	.0037

The variance (or its square root) then functions as the appropriateness index for the examinee. If the item in the model presented above is also assumed to be fixed, then the random variables in it are α, θ, and γ.

These random variables are assumed to be sampled only once for each test administration rather than each time an additional item is administered, as in the Gaussian model. Their underlying sampling distributions are across testing occasions rather than item administrations. Interest is centered only on their current values, hence no assumptions regarding distributional forms need be made. The accuracy of the estimates of α, θ, and γ is dependent on the adequacy of sampling from a homogeneous universe of possible test items, whereas the Gaussian model considers the test as a fixed entity, with no sampling having taken place.

Ultimately, the choice of a test model must depend on the plausibility of its assumptions regard-

ing statistical sampling processes, which can be assessed by goodness-of-fit criteria that take into account model complexity. Considerations involving necessary sample sizes and test lengths as well as the expense of model fitting may mean that the most plausible models can rarely, if ever, be applied. However, advances in computing technology make the application of models such as the Gaussian model and the generalized logistic model more feasible.

Comparisons With Other Studies

As with any linear or nonlinear model containing a large number of parameters, the possibility of acceptable estimation using a finite sample of data is an initial question that must be addressed by empirical monte carlo techniques, even before its applicability to real data can be examined; such was the purpose of this study. One way in which the results can be put into perspective is by comparing them with findings from similar studies, keeping in mind that any variations must be interpreted cautiously in light of the differences in models, sample sizes, test lengths, parameter sampling distributions, computational procedures, and convergence criteria.

Hulin, Lissak, and Drasgow (1982) and Ree (1979) investigated the accuracy of simultaneous parameter estimation for the three-parameter logistic model as it is achieved by the LOGIST algorithm (Wood et al., 1976). The correlations between simulated and estimated parameters common to the three-parameter and generalized logistic models for all three studies are displayed in Table 6. Ree (1979) obtained correlations that consistently exceed the values obtained here, perhaps because of differences in sample size, model complexity, and parameter distributions. The difference in model complexity might also explain the higher correlation for $\hat{\theta}$ achieved by Hulin et al. (1982). Their lower correlations for item parameters are somewhat unexpected, but differences in sampling distributions and estimation procedures probably affected the results. Hulin et al. also reported root mean squared difference statistics, but differences in computational method preclude meaningful comparisons with the values obtained here.

Conclusions

Much more work needs to be done before this model can be applied. The next step in its development is to improve the parameter estimation algorithm as much as possible, and then apply it to real data. If the fit to real data with large values of n and N is worse than that provided by the three-parameter logistic model, then the full model with all three person parameters and three item parameters should be discarded.

If this model does give a better fit to real data, then the next logical step would be to test whether the added parameters provide reliable information on individual differences relating to test score appropriateness. A study similar to that carried out by Drasgow (1982) would be ideal. Only if the extra person parameters offer reliable information surpassing or supplementing that provided by simpler techniques can the use of this model be justified. Even if the use of this general model is unsupported after further study, it is still possible that one or more of its submodels, for example the submodel that assumes the person asymptotes are

Table 6
Correlations Between True and Estimated Parameters

Study	N	n	a	b	c	θ
Present study	1000	111	.781	.967	.178	.873
Hulin et al. (1982)	1000	60	.543	.939	--	.898
Ree (1979)	2000	80	.827	.975	.379	.965

all equal to 0 or to some other constant, might provide a good model for some measurement applications.

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2,* 451–462.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement, 28,* 105–113.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6,* 297–308.

Drasgow, F., Levine, M. V., & Williams, E. A. (1984). *Appropriateness measurement with polychotomous item response models and standardized indices* (Measurement Series 84–1). Champaign IL: University of Illinois.

Fletcher, R. (1972). *FORTRAN subroutines for minimization by quasi-Newton methods* (Report R7125). Harwell, England: Atomic Energy Research Establishment.

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

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6,* 249–260.

IMSL Library (1982, 9th ed.). Houston TX: International Mathematical and Statistical Libraries.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42–56.

Levine, M. V., & Rubin, D. B. (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., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

Lumsden, J. (1977). Person reliability. *Applied Psychological Measurement, 1,* 477–482.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31,* 19–26.

Lumsden, J. (1980). Variations on a theme by Thurstone. *Applied Psychological Measurement, 4,* 1–7.

Ree, M. J. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3,* 371–385.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81–96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215–231.

Tatsuoka, M. M. (1979). Recent psychometric developments in Japan: Engineers tackle educational measurement problems. *ONR—Tokyo Science Bulletin, 4,* 1–7.

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

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models* (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology. (NTIS No. AD-A080 933)

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 83–108). New York: Academic Press.

Vale, C. D., & Weiss, D. J. (1975). *A study of computer-administered stradaptive ability testing* (Research Report 75–4). Minneapolis: University of Minnesota, Department of Psychology. (NTIS No. AD-A018 758)

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73–3). Minneapolis: University of Minnesota, Department of Psychology. (NTIS No. AD-768 376)

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

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Acknowledgments

This article is based on the master's thesis of the first author, conducted while a predoctoral trainee in the Quantitative Methods Program of the Department of Psychology, University of Illinois at Urbana-Champaign. The research was supported in part by the Al-

cohol, Drug Abuse, and Mental Health Administration, National Research Service Award No. MH14257. The authors thank Charles L. Hulin for suggesting some of the data analyses and commenting on earlier drafts of this article, and Michael V. Levine for suggesting a useful parameter estimation strategy.

Author's Address

Send requests for reprints or further information to Nancy L. Strandmark, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign IL 61820, U.S.A.