Distinguishing Among Parametric Item Response Models for Polychotomous Ordered Data

Albert Maydeu-Olivares, Fritz Drasgow, and Alan D. Mead

University of Illinois

Several item response models have been proposed for fitting Likert-type data. Thissen & Steinberg (1986) classified most of these models into difference models and divide-by-total models. Although they have different mathematical forms, divide-by-total and difference models with the same number of parameters seem to provide very similar fit to the data. The ideal observer method was used to compare two models with the same number of parameters--Samejima's (1969) graded response model (a difference model) and Thissen & Steinberg's (1986) extension of Masters' (1982) partial credit model (a divide-by-total model)—to investigate whether difference models or divide-by-total models should be preferred for fitting Likert-type data. The models were found to be very similar under the conditions investigated, which included scale lengths from 5 to 25 items (five-option items were used) and calibration samples of 250 to 3,000. The results suggest that both models fit approximately equally well in most practical applications. *Index terms: graded response model, IRT, Likert scales, partial credit model, polychotomous models, psychometrics.*

Although several item response theory (IRT) models for ordered polychotomous items have been proposed (e.g., Andrich, 1978a, 1978b; Masters, 1982; Samejima, 1969), relatively little is known about how well each of these models fits Likert scale items (see Dodd, 1984; Koch, 1983; Maydeu-Olivares, 1991, 1993; Reise & Yu, 1990). In this paper, some of the most commonly used IRT models for ordered polychotomous data are described using Thissen & Steinberg's (1986) taxonomy as a framework. The existing literature is reviewed on how well these models fit actual and simulated data. Finally, an empirical study is reported that investigated the degree to which two models with different mathematical forms are distinguishable in their predictions under identical experimental conditions (e.g., number of parameters per model, estimation method, size of calibration sample, number of items, number of options per item).

Models for Ordered Polychotomous Items

Thissen & Steinberg (1986) proposed a taxonomy for unidimensional parametric item response models that consists of binary models and complex models. Binary models are "... models for free-response binary test items" (p. 569) and can be used in the analysis of dichotomously scored responses; they also are used as the building blocks for complex models. The normal ogive model, the one- and two-parameter logistic models, and various spline models (Ramsay, 1988; Winsberg, Thissen, & Wainer, 1984) are examples of binary models.

Thissen & Steinberg's (1986) taxonomy for complex models consists of (1) difference models; (2) divide-by-total models; (3) left-side-added models; and (4) left-side-added divide-by-total models. The assumption that there is no guessing or similar psychological phenomenon underlying ordered polychotomous responses (i.e., Likert scale items) suggests option response functions (ORFs) with zero lower asymptotes. Because of that assumption, the difference and the divide-by-total categories of Thissen and Steinberg's taxonomy are of primary interest.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 18, No. 3, September 1994, pp. 245–256

0146-6216/94/030245-12$1.85

Samejima's Graded Response Model

The best known example of a difference model is Samejima's (1969) graded response model (GRM), which has its origins in Thurstone's law of categorical judgment (see Thurstone, 1959). The GRM divides the m categories of an item into $m-1$ blocks. Each of these blocks can be thought of as an item response function modeling a dichotomous process: selecting options 1, 2, ..., k versus option $k+1$, $k+2$, ..., m. Each of these blocks (e.g., option 1 versus option 2 or above, option 2 or below versus option 3 or above) is denoted by the probability $P^*(u_i \geq k+1|\theta = t)$. In general, any binary model can be used to model this probability. Samejima (1969) suggested using normal ogives or two-parameter logistic functions for this purpose. Logistic functions were used here to model the conditional probabilities for the $m-1$ blocks. Consequently $P^*(u_i \geq k+1|\theta = t)$ can be expressed as:

$$P^*(u_i \geq k+1|\theta = t) = \frac{1}{1+\exp\left[-a_i\left(t-b_{i,k+1}\right)\right]}, \tag{1}$$

where

u_i denotes the polychotomously scored response to item i,
θ is the latent trait random variable,
t denotes a specific value of θ,
a_i is the single discrimination parameter of the item, and
$b_{i,k+1}$ is the difficulty parameter for selecting option $k+1$ or above.

In the GRM, the probability of selecting option k, given the latent trait $\theta = t$, is computed as

$$P(u_i = k|\theta = t) = P^*(u_i \geq k|\theta = t) - P^*(u_i \geq k+1|\theta = t). \tag{2}$$

An important characteristic of the GRM is that a must be equal for all options of an item. That is, there must be a common a_i parameter for all P^* within an item. Otherwise, the $P^*(u_i \geq k+1|\theta = t)$ will cross and negative probabilities would be obtained for some ORFs, $P(u_i = k|\theta = t)$. The GRM uses a single a parameter for all options of a particular item but allows different a parameters across items. Takane & de Leeuw (1987) showed that the GRM with P^* modeled by normal ogives is equivalent to factor analysis models for categorical variables described by Christoffersson (1975) and Muthén (1983, 1984).

Thissen and Steinberg's Extension of Masters' Partial Credit Model

Masters' (1982) partial credit model (PCM) is a divide-by-total model. Masters proposed using one-parameter logistic functions in each of the $m-1$ blocks in order to obtain a model with Rasch properties (see Rasch, 1960, 1961). Thus, in the PCM, a one-parameter logistic function is used to model the probability that an individual selects option 2 given that he/she has selected option 1 or 2. A one-parameter logistic function also is used to model the probability that an individual selects option 3 given that he/she has selected options 2 or 3, and so forth. One-parameter logistic functions are fit to all $m-1$ dichotomies obtained from subsamples who select option $k+1$ rather than k. Masters showed that using one-parameter logistic functions produced sufficient statistics for the θ parameter and b parameters (i.e., "parameter separability;" therefore, "specifically objective" measurement was achieved).

The conditional probabilities of responses for the $m-1$ subpopulations are denoted by $P^o(u_i = k+1|\theta = t, u_i = k \text{ or } k+1)$. For increased flexibility, Thissen & Steinberg (1986) suggested using two-parameter logistic functions to model these conditional probabilities:

$$P^o(u_i = k+1|\theta = t, u_i = k \text{ or } u_i = k+1) = \frac{1}{1+\exp\left[-a_{i,k+1}\left(t-b_{i,k+1}\right)\right]}. \tag{3}$$

Note that the introduction of a category a parameter increases flexibility in modeling conditional probabilities, but destroys the parameter separability that is central to Rasch models. Andrich (1978a) also considered a model with varying a parameters for the conditional probabilities given in Equation 3 (see his Equations 8a and 8b), but Rasch model considerations led him to impose restrictions on the a parameters.

In Thissen & Steinberg's (1986) model,

$$P^{o}(u_i = k+1 | \theta = t, u_i = k \text{ or } u_i = k+1) = \frac{P(u_i = k+1 | \theta = t)}{P(u_i = k | \theta = t) + P(u_i = k+1 | \theta = t)} . \quad (4)$$

Solving for $P(u_i = k+1 | \theta = t)$ in the system of $m-1$ equations of the form given by Equation 4 with the restriction that

$$\sum_{k=1}^{m} P(u_i = k | \theta = t) = 1 \quad (5)$$

results in

$$P(u_i = k | \theta = t) = \frac{\exp\left[\sum_{k'=1}^{k} a_{i,k'}(t - b_{i,k'})\right]}{\sum_{h=1}^{m} \exp\left[\sum_{k'=1}^{h} a_{i,k'}(t - b_{i,k'})\right]} \quad (6)$$

for $k = 1, 2, ..., m$, with the convention that $a_{i1} = 0$ and $b_{i1} = 0$. Note that the denominator in Equation 6 is the sum of the numerators for the m option probabilities, which is the defining characteristic of Thissen & Steinberg's (1986) "divide-by-total" models.

With the restriction that a is constrained to be equal across options $k = 1, ..., m$ within an item, but not across items, Thissen & Steinberg's (1986) model has the same number of parameters as the GRM. Thissen and Steinberg therefore suggested that the revised model was "... a more direct competitor to Samejima's graded model" (p. 571). The model defined in Equation 6 with the restriction of common discrimination within items but not across items is an extension of Masters' (1982) PCM and will be referred to as Thissen and Steinberg's ordinal model (TSOM).

Thissen & Steinberg (1986) showed that the TSOM and models described by Masters (1982), Andrich (1978a, 1982, 1985), and Masters & Wright (1984), are constrained versions of Bock's (1972) nominal model. Thus, Bock's nominal model is a more flexible model for polychotomous data than any of the above-mentioned constrained versions. The constrained models impose structure on the ORFs, which seems reasonable given the task of responding to Likert scale items. Furthermore, some of the constrained models have Rasch properties—sufficient statistics for item and person parameters. Finally, note that the GRM is not a Rasch-type model even when its a parameters are constrained to be equal across items (Masters, 1982).

Selecting an IRT Model for Likert-Type Data

Very few studies have compared how these different models fit actual or simulated data under different calibration procedures, sample sizes, test length, and violations of model assumptions. A much more extensive literature has examined some of these issues in the context of models for dichotomous responses (Drasgow, 1989; Hulin, Lissak, & Drasgow, 1982; Swaminathan & Gifford, 1982, 1985, 1986), and researchers have begun only recently to conduct parallel studies for polychotomous models. However, examining the fit of polychotomous models is more complex than studying the fit of a dichotomous model. For example, many studies have correlated estimated b and a parameters with the simulation parameters for two- and three-parameter logistic models (Swaminathan & Gifford, 1982, 1985, 1986). Polychotomous models can have

several *a* and *b* parameters for each item; consequently, evaluating item parameter recovery is more difficult. Moreover, in the divide-by-total models, the meaning of one item parameter can depend on the values of other parameters for that item; therefore, simply correlating estimated and simulated parameters may be inadequate in a polychotomous context.

Reise & Yu (1990) studied item and person parameter recovery in the GRM as estimated by marginal maximum likelihood using MULTILOG 5 (Thissen, 1988), simulating response patterns to 25-item unidimensional tests composed of five-point scaled items. Under these conditions, they concluded that adequate item parameter recovery was obtained with samples of 500 people.

Dodd (1984) used joint maximum likelihood calibration with actual and artificial data to compare the GRM, a simplified version of the GRM in which all *a*s were constrained to be equal, and the PCM. The differences among the models were evaluated by correlating the item and person parameter estimates across models and by inspecting the test information functions yielded by each model. Both the item and person parameter estimates correlated highly across methods. However, the inspection of relative efficiency information plots revealed that the simplified GRM yielded considerably less information than the other two models.

Despite having fewer parameters, the PCM with actual data yielded more information than the GRM except at very high θ levels ($\theta > 3$) (Dodd, 1984). Dodd also showed that with simulated data, the GRM yielded more information than the PCM throughout the θ continuum, but only noticeably at its extremes ($|\theta| > 1.5$). The high relative efficiency of the PCM around the center of the θ continuum was attributed to the fact that although all items contributed equally to the test information function (all items had the same *a*), their item information functions (IIFs) provided more information at θ levels that fell in the range of the *b* parameter estimates, and were more peaked when the range of an item's estimated *b* parameters was small. In the GRM, the shape of the IIF depends on the *a* parameters. When the *a*s were forced to be equal, Dodd observed that the resulting IIFs were rather flat.

Maydeu-Olivares (1993) assessed how well several IRT models fit the responses of 1,053 people to five Likert-type scales consisting of between five and 20 five-option items. Among the models studied were the GRM, the PCM, the TSOM, and Bock's nominal model. Parameters of all models were estimated using marginal maximum likelihood as implemented in MULTILOG 6 (Thissen, 1991). A normal ogive version of the GRM estimated by generalized least squares as implemented in LISCOMP (Muthén, 1987), and Levine's (1984) nonparametric multilinear formula score model for polychotomous data estimated by marginal maximum likelihood as implemented in ForScore (Williams & Levine, 1993) also were fit to these data. Goodness-of-fit was determined by χ^2 goodness-of-fit statistics computed for each item, for pairs of items, and triples of items within inventories. In addition, fit plots were constructed to compare empirical proportions selecting each option to estimated ORFs.

The inspection of the χ^2 statistics for item pairs and triples revealed that, across all five inventories (1) the full information version of the GRM (as implemented in MULTILOG) slightly outperformed the limited information version of the GRM (as implemented in LISCOMP); (2) the GRM (regardless of estimation method) slightly outperformed the divide-by-total models (the PCM, the TSOM, and Bock's nominal model); (3) among the divide-by-total models, the models with more parameters slightly outperformed the models with fewer parameters (i.e., Bock's model outperformed the TSOM, which in turn outperformed the PCM); (4) Levine's model clearly outperformed all other models. The model with the smallest single-item χ^2 statistics across all five inventories was the GRM as implemented in LISCOMP. Of special relevance is the fact that Bock's nominal model was not able to outperform the GRM with respect to the item pair and item triple fit statistics despite its larger number of parameters. The differences observed when using one parametric model or another were not large, however, and thus it seemed legitimate to ask the question of how large the differences are between these models, and how relevant they are for applications.

Measuring the Difference Between Two Models

The Ideal Observer Index

The ideal observer index (IOI; Levine, Drasgow, Williams, McCusker, & Thomasson, 1992) was used here to study some of the properties of two of the most promising models for Likert scale items—the GRM and the TSOM. The IOI is designed to facilitate comparisons of IRT models that may vary in any of a number of ways. In fact, this method may be applied whenever there are two statistical models for the n items in a scale. More specifically, the IOI can be used if there are two ways of computing the probability of each of the m^n response patterns for n items with m options.

The IOI has its roots in signal detection theory (Green & Swets, 1966) in which the two-alternative forced-choice experiment provides an important means for quantifying the discriminability of two stimuli—Stimulus A and Stimulus B. This experiment consists of a set of trials. On each trial, one stimulus is presented first, and then the other stimulus is presented. The observer knows that there is an equal probability (i.e., .5) of each stimulus being presented first. After the two stimuli have been presented, the observer is asked to decide whether the stimuli were presented in the order AB or BA. If the stimuli are impossible to differentiate, the observer should be correct 50% of the time. As the observer's ability to correctly identify the two stimuli increases, the correct classification rate increases to a maximum of 100%.

The IOI is an extension of the two-alternative, forced-choice experiment to the context of IRT in which two statistical models for response patterns are compared. Here, "Stimulus A" is a randomly sampled response pattern from Model A, and "Stimulus B" is a randomly sampled response pattern from Model B. If the Model A probability is virtually identical to the Model B probability for each of the m^n response patterns, then it is nearly impossible to differentiate between the models, and a correct classification rate close to .5 will be obtained. If the two models have substantially different probabilities for the response patterns, an observer would be able to use this information and achieve a classification rate above .5.

The IOI is based on an ideal observer, namely an observer who uses a most powerful test for differentiating between the two statistical models. Specifically, a response pattern, $\mathbf{u}_1$, is randomly sampled from Model A and another response pattern, $\mathbf{u}_2$, is randomly sampled from Model B. After the response patterns have been sampled, the task is to classify one pattern as the Model A pattern and the other pattern as the Model B pattern. An ideal observer bases this decision on likelihood ratios

$$\lambda(\mathbf{u}_1) = \frac{P_A(\mathbf{u}_1)}{P_B(\mathbf{u}_1)} \tag{7}$$

and

$$\lambda(\mathbf{u}_2) = \frac{P_A(\mathbf{u}_2)}{P_B(\mathbf{u}_2)}. \tag{8}$$

Levine et al. (1992) showed that the most accurate classification is based on the decision rule

> Classify $\mathbf{u}_1$ as a Model A pattern and $\mathbf{u}_2$ as a Model B pattern if $\lambda(\mathbf{u}_1) > \lambda(\mathbf{u}_2)$; otherwise classify $\mathbf{u}_1$ as a Model B pattern and $\mathbf{u}_2$ as a Model A pattern.

The IOI is defined as the correct classification rate of this rule.

The IOI is closely related to Akaike's information criterion (AIC; Akaike, 1987; Bozdogan, 1987; Takane, 1994). The AIC differs from the IOI in that the AIC is a function of the likelihood of a single model fit to the data; the IOI is based on a likelihood ratio of two models, as shown in Equations 7 and 8. However, it may be possible to develop connections between the two approaches to fit by considering the theory underlying the AIC (see Bozdogan, 1987, for a description of this theory). In addition, application of the AIC to the problem of model fit in IRT appears to be an interesting area for future research.

The value of the IOI has a relatively straightforward interpretation. An IOI near 1.0 means that response patterns from the two models are easily differentiated: The ideal observer is almost always correct in classification decisions. Alternatively, an IOI value slightly above .5 means that the models are virtually indistinguishable in that an optimal classification procedure does little better than a classification decision based on a random event (e.g., the toss of a coin). In this case, almost every response pattern must have almost exactly the same probability for both models. Values less than .5 are impossible for an optimal decision maker.

Uses of the IOI

The IOI can be used for a wide variety of comparisons. Levine et al. (1992), for example, used the IOI to gauge how well BILOG (Mislevy & Bock, 1989), LOGIST (Wingersky & Lord, 1982), and ForScore (Williams & Levine, 1993) estimated three-parameter logistic item response functions. They found IOI values ranging from approximately .55 to .65 when calibration samples of $N = 3{,}000$ were used. A study in progress examines how calibration accuracy—as indexed by the IOI—improves as sample size increases. Comparisons of other pairs of models could include the following: a model obtained from an analysis of a sample of majority group members and a model obtained from a minority group, in the analysis of differential item functioning; a model defined by the parameter estimates after j stages of an iterative calibration algorithm and a model defined by the estimates after $j + 1$ stages, to determine whether the iterative calibration procedure converged; and a unidimensional model versus a multidimensional model for a given dataset.

In this study, the IOI was used to compare a model obtained by one particular mathematical representation of ORFs to a model defined by an alternative mathematical formulation. Using simulation data allowed multiple replications per cell of the experimental design, and should increase the sensitivity of the comparisons. Use of the IOI as the measure of model fit was intended to improve the power and accuracy of model comparisons.

A second issue examined in this paper concerned the minimum sample sizes needed to analyze Likert scale items with an appropriate IRT model for varying scale lengths. In contrast to ability tests, in which it is often possible to obtain large datasets from archival sources, attitude and self-report inventory data ordinarily must be collected by the researcher, in some cases through individual interviews.

Method

Simulation Design

Three scale lengths ($n = 5$, 15, and 25 items) and four sample sizes ($N = 250$, 500, 1,000, and 3,000 simulated respondents) were examined. For each of the 24 sets of items (two models × three ns × four Ns), three samples were generated independently using a standard normal distribution. 25 items and 3,000 respondents was selected as the largest combination of N and n because: (1) unidimensional attitude scales and self-report inventories are not usually longer than 25 items, and (2) a sample of 3,000 is very large for an attitude scale or a self-report inventory. Small Ns and ns also were selected because it was expected that these conditions would show the largest calibration errors. Moreover, the purpose was to study calibration errors under circumstances that are likely to be encountered in practice.

Simulation Item Parameters

To increase realism, simulation item parameters were based on actual data. Specifically, the set of 25 item parameters used in this study was obtained by merging the parameters of the Positive Problem Orientation and Rational Problem Solving scales of the Social Problem Solving Inventory–Revised (D'Zurilla & Maydeu-Olivares, 1993). The five-point Likert type items of this inventory had been calibrated in a previous study (Maydeu-Olivares, 1993) with MULTILOG 6 (Thissen, 1991) using the GRM and the TSOM. These estimates

subsequently were used as simulation parameters.

The 5-, 15-, and 25-item parameter sets used were hierarchically nested. That is, the 15 items were a subset of the 25 items described above, and the 5 items were a subset of the 15 items. The items to be included in the 5- and 15-item subsets were selected (1) by distributing the 25 items into five, 5-item clusters so that each cluster would contain items with similar IIFs; and (2) by sampling without replacement one item from each cluster to obtain the 5-item set, and three items from each cluster to obtain the 15-item set.

Data Analysis

The 72 resulting datasets were analyzed by both the GRM and the TSOM using MULTILOG 6 (Thissen, 1991). Thus, data generated by the GRM, for example, were calibrated with both the model used to generate the data and the TSOM. Consequently, it could be determined how well the GRM and the TSOM reproduced probabilities of response patterns from data generated by their own model, and how well these models reproduced response pattern probabilities from data generated by the alternative model. The analysis was concerned with determining (1) if one model was better than the other in describing response probabilities when it was used to analyze data that satisfied its assumptions, and (2) if one model was better in describing response probabilities when item parameters were estimated from data generated by the alternative model. MULTILOG default specifications were used (i.e., marginal maximum likelihood calibration assuming a standard normal θ distribution), except for the maximum number of EM cycles, which was incremented to 100 to ensure convergence.

Following each MULTILOG run, the "estimated" model (i.e., the statistical model defined by the estimated parameters and the functional form of the ORFs used during parameter calibration) was compared to the simulation model (i.e., the simulation item parameters and the parametric form of the ORFs used to generate the data input to MULTILOG) by means of the IOI.

The Criterion for Evaluating Calibration Accuracy

The IOI method indicated the extent to which it was possible to differentiate response patterns generated by the simulation model from response patterns generated by the estimated model. If it was difficult to differentiate response patterns (i.e., if the IOI correct classification rate was only slightly above .50), then the estimated model was "close" to the simulation model. In such a case, use of the estimated model in place of the true model would be justified. Alternatively, if the IOI classification rate was large, then the estimated model is a poor approximation of the simulation model and the two models should not be used interchangeably.

In the next step of the analysis, IOI was computed by the following process (see Levine et al., 1992, for additional details). First, random samples of the specified number of response patterns were generated using the estimated model and the same number of response patterns were generated using the simulation model. The likelihood ratios in Equations 7 and 8 then were computed for each response pattern and a likelihood ratio for a response pattern generated by the estimated model (Model A) was compared to a likelihood ratio for a response pattern from the simulation model (Model B); Levine et al.'s decision rule for classifying patterns then was applied. Because simulated data were used, it could be determined whether the classification decision was correct. Aggregating over the Model A and Model B response patterns provided a sample estimate of the IOI correct classification rate.

Because some sampling fluctuation in estimates of IOI classification rates was expected due to the particular random sample of response patterns, 10 replications of the above process were performed each time two models were compared. Then the 10 sample estimates of the IOI classification rate were averaged to control for the variability of random samples. The estimated standard error of the IOI for a fixed set of item parameter estimates across 10 replications was generally between .001 and .003.

Results

The means of the IOI and its standard errors from the three replications performed on each on the 48 different experimental conditions (two simulation models × two estimated models × three *n*s × four *N*s) are presented in Table 1. The IOI values in Table 1 range from approximately .527 to approximately .714, indicating that in some cases the estimated models were virtually indistinguishable from the simulation model and in other cases the estimated models were substantially different from the simulation model. The IOI values increased as *n* increased for calibration samples of a given size. This occurred because a longer scale provides more information to a decision maker, and thereby facilitates differentiation between true and estimated models. Thus, to use the IOI as a measure of calibration accuracy, rather than a measure of classification accuracy, some transformation of the IOI to a constant scale length is needed.

Table 1
Means and Estimated Standard Errors (SEs) of IOI Based on Scale Calibrations by MULTILOG for $n = 5$, 15, and 25 Items for Data Simulated by the GRM and TSOM and Estimated by the GRM (Graded) and TSOM (Ordinal), for $N = 250$, 500, 1,000, and 3,000

	Data Simulated by GRM						Data Simulated by TSOM					
N and	Graded			Ordinal			Graded			Ordinal		
Statistic	5	15	25	5	15	25	5	15	25	5	15	25
$N = 250$												
Mean	.576	.653	.687	.579	.668	.708	.588	.678	.714	.583	.664	.699
SE	.002	.004	.004	.002	.005	.008	.003	.007	.012	.003	.009	.013
$N = 500$												
Mean	.559	.610	.643	.565	.631	.671	.570	.633	.677	.564	.620	.655
SE	.006	.003	.003	.005	.005	.003	.002	.002	.004	.002	.003	.006
$N = 1{,}000$												
Mean	.540	.571	.615	.548	.600	.647	.555	.602	.659	.547	.577	.629
SE	.004	.005	.008	.004	.003	.007	.003	.004	.006	.004	.005	.010
$N = 3{,}000$												
Mean	.527	.550	.573	.536	.585	.615	.544	.600	.634	.543	.572	.594
SE	.001	0.000	.003	.003	.001	.002	.003	.003	.003	.004	.004	.004

The effects of *N* are also clear in Table 1: The IOI decreased as *N* increased. This means that the pattern probabilities computed from estimated item parameters became more similar to the simulation model pattern probabilities as larger samples were used to estimate item parameters. Thus, differentiation between simulation model response patterns and estimated model response patterns became more difficult as the size of the calibration sample increased. Table 1 also shows that a model fits data generated by itself somewhat better than data generated by the alternative model.

Regression Analyses

Method. To provide a quantitative model for the trends observed in Table 1, a series of regression analyses were conducted. In these analyses, the IOI logit—ln[IOI/(1 – IOI)]—was used as the dependent variable, and logarithmic transformations of calibration sample size, number of items, and their interaction—ln(*N*), ln(*n*), and ln(*N*) × ln(*n*)—were used as independent variables.

Three types of regression analyses were conducted.

1. A separate regression was run for each combination of the model used to generate data and the model used for test calibration. This resulted in four models—(1) the data were simulated by the GRM and estimated by the GRM, (2) the data were simulated by the GRM and estimated by the TSOM, (3) the data were simulated by the TSOM and estimated by the TSOM, and (4) the data were simulated by the TSOM

and estimated by the GRM. The regression model was

$$\ln[\mathrm{IOI}/(1-\mathrm{IOI})] = b_0 + b_1\ln(n) + b_2\ln(N) + b_3\ln(n)\times\ln(N). \quad (9)$$

2. A regression was run for each model used to generate data, which included a dummy variable for the model used for estimation (i.e., two models). The regression model was

$$\ln[\mathrm{IOI}/(1-\mathrm{IOI})] = b_0 + b_1\ln(n) + b_2\ln(N) + b_3\ln(n)\times\ln(N) + b_4(\text{estimation model}). \quad (10)$$

3. A regression was run for a single overall model, which included dummy variables indicating the model used to simulate data and the model used for estimation. The regression model was

$$\begin{aligned}\ln[\mathrm{IOI}/(1-\mathrm{IOI})] = {} & b_0 + b_1\ln(n) + b_2\ln(N) + b_3\ln(n)\times\ln(N) + b_4(\text{estimation model}) \\ & + b_5(\text{data simulation model}) + b_6(\text{estimation model}\times\text{data simulation model}).\end{aligned} \quad (11)$$

All regression models were constructed by a hierarchical procedure. That is, in the first stage all independent variables were entered in the equation. Then interaction terms were examined to determine whether they could be omitted from the regression equation without significantly reducing the squared multiple correlation (R^2). Finally, main effects were examined to determine whether they could be omitted.

A related set of significance tests was performed for each regression equation, and consequently the Bonferroni correction was applied to maintain an overall α level of .05 per equation. Thus, because the separate regression equations for each combination of data simulation model and estimation model contained three terms, $\alpha = .05/3 = .0167$ was used for the individual significance tests. Significance tests for Regressions 2 and 3 (Equations 10 and 11, respectively) used $\alpha = .05/4 = .0125$ and $.05/6 = .0083$, respectively.

Results. Table 2 presents the results of the regression analyses. Only those parameter estimates significant at an overall $\alpha = .05$ are reported. All regression equations resulted in $R^2 > .91$. All regression weights reported in this table were very large compared to their standard errors. As shown in Table 2 for Regression 1, the regression models for the logit transformation of IOI values obtained by fitting a model to itself (i.e., the model used to estimate item parameters for a scale was the same as the model used to simulate the data) included a test length main effect and a sample size × test length interaction.

Fitting separate regression equations of the form in Equation 10 (Regression 2) for the data generated according to the GRM and the TSOM allowed for a comparison of whether a model fit its own data (i.e., generated according to its assumptions) better than data generated according to the alternative model. In both cases, the regression coefficient for the estimation model dummy variable was significant ($t = -.044/.006 = -7.3$, $p < .0001$ for the GRM and $t = .037/.008 = 4.6$, $p < .0001$ for the TSOM). Thus, both models better fit data generated according to their assumptions than data generated by the alternative model.

When a single overall model was fit (Regression 3), the nonsignificant effects included the estimation model main effect ($|t| = .723$, $p = .471$) and the N main effect ($|t| = 1.223$, $p = .223$). All other terms were highly significant ($p < .0001$). The strong interaction obtained between the estimation and simulation models reflects the fact that both models fit themselves better than the alternative parametric model. The simulation model main effect found to be significant was due to the fact that a slightly better fit was obtained for samples generated by the GRM.

Discussion

Using logarithmic transformations of scale length and sample size, the logit transformations of the IOI were predicted very accurately. The regression equations obtained can be used to interpolate IOI values within the range of Ns and ns studied here. For example, the IOI for the GRM with 15 items and a calibration sample of 1,500 respondents could be estimated for a set of items similar to the one studied.

Similar regression equations were found to be appropriate under all conditions. Specifically, the regres-

Table 2
Significant Regression Coefficients (b) and their SEs from Regression Models Predicting IOI Values From Estimation Sample Size and Test Length for Data Generated by the GRM and TSOM and Estimated by the GRM (Graded) and TSOM (Ordinal) and by an Overall Model

Model, Coefficient, and R^2	GRM				TSOM				Overall Model	
	Graded		Ordinal		Graded		Ordinal			
	b	SE	b	SE	b	SE	b	SE	b	SE
1. Combinations of Simulation and Estimation Models										
b_0	−.127	.030	−.206	.029	−.197	.040	−.121	.042		
b_1	.599	.024	.605	.024	.582	.032	.585	.035		
b_2										
b_3	−.060	.003	−.051	.003	−.046	.004	−.055	.005		
R^2	.95		.96		.93		.91			
2. Dummy Variable for the Estimation Model										
b_0	−.167	.023			−.160	.030				
b_1	.602	.019			.583	.025				
b_2										
b_3	−.055	.002			−.050	.003				
b_4	−.044	.006			.037	.008				
R^2	.95				.91					
3. Overall Regression Model										
b_0									−.163	.019
b_1									.593	.016
b_2										
b_3									−.040	.002
b_4										
b_5									−.022	.005
b_6									−.040	.005
R^2									.93	

sion equations included a scale length main effect and a scale length × sample size interaction. Thus, the degree of similarity between the simulation model and estimated model depended on the size of the calibration sample through its interaction with scale length. Of course, in a simulation study that held scale length constant a sample size main effect would be expected.

Not surprisingly, both models studied here provided a better fit to data generated by the same parametric model rather than the alternative parametric model. However, neither model was found to be uniformly superior in modeling data generated by its own parametric form, and in modeling data generated by the alternative model. Thus, which model is better could not be determined; each model provided a slightly better fit (in samples used to compute the IOI statistic) when its assumptions were satisfied.

Conclusions

Researchers and practitioners alike are interested in the question of which logistic model should be used given a set of Likert-type items written to assess a psychological construct. This question was studied by considering the degree to which the response probabilities of a model fit to a dataset approximated the response probabilities of a simulation model. The ideal observer method and its computational procedures provide one approach (Levine et al., 1992). The IOI gives the accuracy of a most powerful statistical test for classifying response patterns in the two-alternative, forced-choice experiment.

Thissen & Steinberg (1986) presented a taxonomy of existing unidimensional parametric IRT models that helped to delimit the question. According to their taxonomy, there are two sets of IRT models that are most suitable for fitting Likert-type data: difference models and divide-by-total models. The IOI was used to compare two parametric models for Likert-type data with identical numbers of parameters—the GRM and

the TSOM—under identical estimation conditions (marginal maximum likelihood). Under the conditions used in this study, these models proved to be very similar, thus suggesting that either model would be equally appropriate in most practical applications.

References

Akaike, H. (1987). Factor analysis and the AIC. *Psychometrika, 52,* 317–332.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 38,* 123–140.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581–594.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105–113.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco: Jossey-Bass.

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

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

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent models (Doctoral dissertation, University of Texas at Austin). *Dissertation Abstracts International, 45,* 2074A.

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

D'Zurilla, T. J., & Maydeu-Olivares, A. (1993). *A revision of the Social Problem-Solving Inventory based on factor-analytic methods: An integration of theory and data.* Manuscript submitted for publication.

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

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.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15–32.

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

Levine, M. V., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. L. (1992). Distinguishing between item response theory models. *Applied Psychological Measurement, 16,* 261–278.

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

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

Maydeu-Olivares, A. (1991). *Un modelo de ecuaciones estructurales de la resolución de problemas sociales. Efectos de la aplicación de la teoría de respuesta a los items* [A structural equations model of social problem solving: Effects of the application of item response theory]. Unpublished doctoral dissertation, University of Barcelona.

Maydeu-Olivares, A. (1993). *Fitting unidimensional item response models to actual Likert-type inventories.* Unpublished manuscript, University of Illinois, Department of Psychology, Champaign.

Mislevy, R. J., & Bock, R. D. (1989). *PC-BILOG 3: Item analysis and test scoring with binary scoring models* [Computer program]. Mooresville IN: Scientific Software.

Muthén, B. (1983). Latent variable structural modeling with categorical data. *Journal of Econometrics, 22,* 43–65.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations using a comprehensive measurement model* [Computer program]. Mooresville IN: Scientific Software.

Ramsay, J. O. (1988). Monotone regression splines in action (with discussion). *Statistical Science, 3,* 425–461.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danmarks Paedogogiske Institut (Expanded edition, University of Chicago Press, 1980).

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, IV* (pp. 321–334). Berkeley: University of California Press.

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

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

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

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

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

Takane, Y. (1994). A review of applications of AIC in psychometrics. In H. Bozdogan (Ed.), *Proceedings of US/Japan modeling conference* (pp. 379–403). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Thissen, D. (1988). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 5.1) [Computer program]. Mooresville IN: Scientific Software.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 6) [Computer program]. Mooresville IN: Scientific Software.

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

Thurstone, L. L. (1959). *The measurement of values.* Chicago: University of Chicago Press.

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

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

Winsberg, S., Thissen, D., & Wainer, H. (1984). *Fitting item characteristic curves with spline functions* (Technical Report No. 84-52). Princeton NJ: Educational Testing Service.

Acknowledgments

This research was supported in part by Contracts No. N00014-89-K-059 and N00014-90-J-1958 from the Office of Naval Technology, and by Office of Naval Research Contracts No. N00014-86K-0482, NR 442-1546, Michael Levine, principal investigator. The participation of the first author in this research was supported by a Fulbright La Caixa Scholarship, and by a Postdoctoral Scholarship from the Ministry of Education and Science of Spain. Parts of this paper were presented at the 57th Annual Meeting of the Psychometric Society in Columbus OH. The authors thank Bruce Williams for programming assistance, and Ulf Böckenholt and especially Michael Levine for their insightful comments on an earlier draft of this paper.

Author's Address

Send requests for reprints or further information to Albert Maydeu-Olivares, Dto. Estadistica y Econometrica, Universidad Carlos III, C/Madrid 126, 28903 Getafe (Madrid), Spain. Internet: amaydeu@est-econ.uc3m.es.