A Unidimensional Item Response Model for Unfolding Responses From a Graded Disagree–Agree Response Scale

James S. Roberts, Educational Testing Service

James E. Laughlin, University of South Carolina

Binary or graded *disagree–agree* responses to attitude items are often collected for the purpose of attitude measurement. Although such data are sometimes analyzed with cumulative measurement models, recent studies suggest that unfolding models are more appropriate (Roberts, 1995; van Schuur & Kiers, 1994). Advances in item response theory (IRT) have led to the development of several parametric unfolding models for binary data (Andrich, 1988; Andrich & Luo, 1993; Hoijtink, 1991); however, IRT models for unfolding graded responses have not been proposed. A parametric IRT model for unfolding either binary or graded responses is developed here. The graded unfolding model (GUM) is a generalization of Andrich & Luo's hyperbolic cosine model for binary data. A joint maximum likelihood procedure was implemented to estimate GUM parameters, and a subsequent recovery simulation showed that reasonably accurate estimates could be obtained with minimal data demands (e.g., as few as 100 respondents and 15 to 20 six-category items). The applicability of the GUM to common attitude testing situations is illustrated with real data on student attitudes toward capital punishment. *Index terms: attitude measurement, graded unfolding model, hyperbolic cosine model, ideal point process, item response theory, Likert scale, Thurstone scale, unfolding model, unidimensional scaling.*

Using *disagree–agree* responses for measuring attitudes has a long history within psychology. The practice can be traced to Thurstone's (1928, 1931) classic approach in which persons provide binary disagree–agree responses to a set of prescaled attitude statements, and then the responses are used to develop estimates of each person's attitude. Use of disagree–agree responses was bolstered further by the popularity of Likert's (1932) attitude measurement procedure in which persons respond to a set of attitude items using a graded scale of agreement (e.g., *strongly disagree*, *disagree*, *agree*, *strongly agree*), and the graded responses are subsequently used to develop a summated attitude score for each person.

Methods used to analyze disagree–agree responses (binary or graded) to attitude items generally follow one of two perspectives about the response process. The first perspective suggests that disagree–agree responses result from an ideal point process (Coombs, 1964) in which a person agrees with an attitude item to the extent that the item content satisfactorily represents the person's own opinion. From this viewpoint, disagree–agree responses are best analyzed with some type of unfolding model that implements a single-peaked response function.

Thurstone's (1928, 1931) attitude measurement procedure is an example of the ideal point perspective because persons are assumed to have attitudes that are similar to the items they endorse. More contemporary examples include parametric item response models such as the squared simple logistic model (Andrich, 1988), the PARELLA model (Hoijtink, 1990, 1991), and the hyperbolic cosine model (Andrich & Luo, 1993). Nonparametric item response models also have been proposed for data that result from an ideal point response process (Cliff, Collins, Zatkin, Gallipeau, & McCormick, 1988; van Schuur, 1984), but

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 231–255
© Copyright 1996 Applied Psychological Measurement Inc.
0146-6216/96/030231-25$2.50

these nonparametric models do not yield measures that are invariant to either the items or the persons sampled in a particular application. Parametric item response models, in contrast, yield invariant measures, provided that the model in question is appropriate for the data (Hoijtink, 1990).

The second perspective implies that disagree–agree responses are the result of a dominance process (Coombs, 1964) in which a person agrees with a positively worded attitude statement to the extent that the person's own opinion is more positive than the sentiment expressed in the statement. (Conversely, a person agrees with a negatively worded statement to the extent that the person's opinion is more negative than the sentiment reflected by the statement.) According to this viewpoint, disagree–agree responses are best analyzed using some type of cumulative model that implements a monotonic response function. The Likert (1932) and Guttman (1950) approaches to attitude measurement illustrate classical techniques that are consistent with the dominance perspective. The application of cumulative item response models [e.g., the one-parameter (Rasch, 1960/80), two-parameter (Birnbaum, 1968), and three-parameter (Lord, 1980) logistic models, the graded response model (Samejima, 1969), the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), and the generalized partial credit model (Muraki, 1992)] to disagree–agree responses provides a more contemporary illustration of the dominance perspective.

Several researchers have argued that disagree–agree responses generally are more consistent with the ideal point perspective than the dominance perspective (Roberts, 1995; Roberts, Laughlin, & Wedell, 1996; van Schuur & Kiers, 1994). This argument implies that attitude measures based on disagree–agree responses are more appropriately developed from unfolding models than from cumulative models. Moreover, Roberts (1995) showed that cumulative models can yield attitude measures that are nonmonotonically related to the latent trait when such models are applied to responses from an ideal point process. Specifically, persons with the most extreme attitudes may receive scores that are indicative of more moderate attitudinal positions.

Although unfolding models appear most appropriate for disagree–agree data, the application of unfolding item response models to attitude measurement remains somewhat problematic because such models allow only for binary disagree–agree responses. Some researchers, however, have documented gains in precision when polytomous responses, as opposed to binary responses, are used to derive measurements from cumulative item response models (Bock, 1972; Donoghue, 1994; Thissen, 1976). Roberts (1995) suggested that similar benefits can be achieved with polytomous unfolding models. Furthermore, researchers often collect graded responses to attitude statements using the traditional Likert (1932) method, and these practitioners would be forced to dichotomize their data and risk losing valuable information before using any of the contemporary unfolding item response models.

An unfolding item response model is proposed here for disagree–agree responses that result from either a binary or graded scale. The model, called the graded unfolding model (GUM), is a generalization of Andrich & Luo's (1993) hyperbolic cosine model for binary data. The GUM incorporates a single peaked response function; therefore, it is applicable in situations in which responses are presumably generated from an ideal point process. Additionally, because the GUM allows for graded responses, there is no need to dichotomize polytomous data prior to estimating model parameters; thus, there is no corresponding loss in the precision of person estimates.

The Graded Unfolding Model

The GUM is developed from four basic premises about the response process. The first premise is that when a person is asked to express his or her agreement with an attitude statement/item, the person tends to agree with the item to the extent that it is located close to his or her own position on a unidimensional latent attitude continuum. In this context, the degree to which the sentiment of an item reflects the opinion of a person is given by the proximity of the person to the item on the attitude continuum. If δ_i denotes the position of the *i*th item on the continuum and θ_j denotes the location of the *j*th person on the continuum, then the person is more likely to agree with the item to the extent that the distance between θ_j and δ_i approaches 0. This is simply a restatement of the fundamental characteristic of an ideal point process (Coombs, 1964).

The second premise of the GUM is that a person may respond in a given response category for either of two distinct reasons. For example, consider a person with a neutral attitude toward capital punishment. This person might strongly disagree with an item that portrays the practice of capital punishment in either a very negative or very positive way. If the item is located far below the person's position on the attitude continuum (i.e., the item's content is much more negative than the person's attitude), then the person *strongly disagrees from above* the item. In contrast, if the item is located far above the person's position (i.e., the item's content is much more positive than the person's attitude), then the person *strongly disagrees from below* the item. Hence, there are two possible subjective responses—*strongly disagree from above* and *strongly disagree from below*—associated with the single observable response of *strongly disagree*. The GUM postulates two subjective responses for each observable response on a rating scale.

The third premise behind the GUM is that subjective responses to attitude statements follow a cumulative item response model (e.g., Andrich & Luo, 1993). In this paper, it is assumed that subjective responses follow Andrich's (1978) rating scale model, but other cumulative models could also be used. (A rating scale model seems particularly appropriate given that the same response scale is typically used for all items on a traditional Likert or Thurstone attitude questionnaire.) The rating scale model for subjective responses is defined as

$$P\left(Y_i = y \middle| \theta_j\right) = \frac{\exp\left[y\left(\theta_j - \delta_i\right) - \sum_{k=0}^{y} \tau_k\right]}{\sum_{w=0}^{M} \exp\left[w\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w} \tau_k\right]}, \tag{1}$$

subject to the constraint that

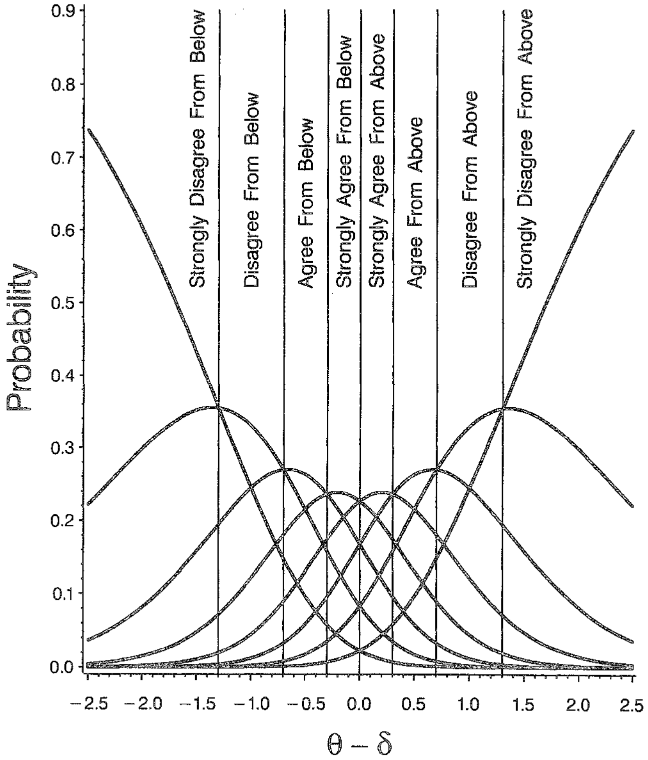
$$\sum_{k=0}^{M} \tau_k = 0, \tag{2}$$

where

Y_i is a subjective response to attitude statement i;
$y = 0, 1, 2, \ldots, M$; $y = 0$ corresponds to the strongest level of disagreement from below the item, and $y = M$ corresponds to the strongest level of disagreement from above the item (see Figure 1);
θ_j is the location of person j on the attitude continuum;
δ_i is the location of attitude statement i on the attitude continuum;
τ_k is the relative location of the kth subjective response category threshold on the attitude continuum (relative to a given item); and
M is the number of subjective response categories minus 1.

The rating scale model is a divide-by-total model (Thissen & Steinberg, 1986). Therefore, the denominator of Equation 1 is simply the sum of all possible numerators that may arise. Note that the value of τ_0 is arbitrarily defined to be 0 in Equation 1, but it could be set equal to any constant without affecting the resulting probabilities (Muraki, 1992). The model is illustrated in Figure 1 for a hypothetical item with four observable response categories—*strongly disagree*, *disagree*, *agree*, and *strongly agree*. The abscissa of Figure 1 represents the attitude continuum, and it is scaled in units of signed distance between a person's attitude position and the location of the item (i.e., $\theta_j - \delta_i$). The ordinate indexes the probability that a person's subjective response will fall in one of the eight possible subjective response categories. [There are eight subjective

Figure 1
Subjective Response Category PFs for a Hypothetical Item
(Subjective Response Category Thresholds Are Located at −1.3, −.7, −.3, 0.0, .3, .7, and 1.3)



response categories and associated probability functions (PFs): A person may respond in any of the four observable response categories because his or her attitudinal position is either above or below the location of the item.] The seven vertical lines designate the locations where successive subjective response category PFs intersect. These locations are the subjective response category thresholds (the τs). In this example, the seven subjective response category thresholds are successively ordered on the latent attitude continuum. Therefore, these thresholds divide the latent continuum into eight intervals in which a different subjective response is most likely. The most likely subjective response within each interval is labeled in Figure 1.

Equation 1 defines an item response model in terms of subjective responses. However, the model must ultimately be defined in terms of the observable response categories associated with the graded agreement scale. Recall that each observable response category is associated with two possible subjective responses (i.e., one from below the item and one from above the item). Moreover, the two subjective responses corresponding to a given observable response category are mutually exclusive. Therefore, the probability that a person will respond using a particular observable category is simply the sum of the probabilities associated with the two corresponding subjective responses:

$$P\left(Z_i = z \middle| \theta_j\right) = P\left(Y_i = z \middle| \theta_j\right) + P\left[Y_i = (M - z) \middle| \theta_j\right], \qquad (3)$$

where

Z_i is an observable response to attitude statement i;

$z = 0, 1, 2, \ldots, C$; $z = 0$ corresponds to the strongest level of disagreement, and $z = C$ refers to the strongest level of agreement; and

C is the number of observable response categories minus 1. Note that $M = 2C + 1$.

The fourth premise behind the GUM is that subjective category thresholds are symmetric about the point $(\theta_j - \delta_i) = 0$, which yields

$$\tau_{(C+1)} = 0 \tag{4}$$

and

$$\tau_z = -\tau_{(M-z+1)}, \quad \text{for } z \neq 0. \tag{5}$$

At a conceptual level, this premise implies that a person is just as likely to agree with an item located at either $-g$ units or $+g$ units from the person's position on the attitude continuum. At an analytical level, this premise leads to the following identity:

$$\sum_{k=0}^{z} \tau_k = \sum_{k=0}^{M-z} \tau_k . \tag{6}$$

Incorporating this identity into Equation 3 yields the formal definition of the GUM:

$$P(Z_i = z | \theta_j) = \frac{\exp\left[z(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_k\right] + \exp\left[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_k\right]}{\sum_{w=0}^{C} \left\{\exp\left[w(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_k\right] + \exp\left[(M - w)(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_k\right]\right\}} . \tag{7}$$

Note that the parameterization used in Equation 7 requires only a single constraint on item parameter values,

$$\sum_{i=1}^{I} \delta_i = 0, \tag{8}$$

along with the notational definition that τ_0 equals 0.

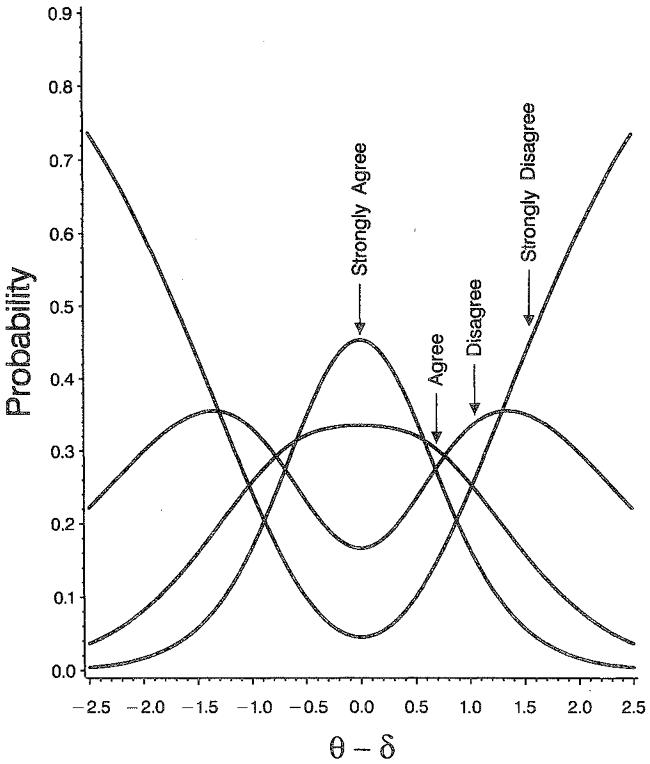
The GUM defines the observable response category PFs associated with the jth person's objective response to the ith item. Figure 2 displays these PFs for the same hypothetical item referenced in Figure 1. Figure 2 shows that there is one PF associated with each observable response available to the person. Each of these PFs is simply the sum of the two corresponding subjective response category PFs previously shown in Figure 1. Note that successive observable response category PFs do not intersect at $\tau_1, \tau_2, \ldots, \tau_C$; therefore, the τ_k parameters lose their simple interpretation at the level of an observable response. In contrast, the substantive meaning of both θ_j and δ_i remains unchanged when moving from subjective responses to observable responses.

The GUM is an unfolding model of the response process. This is easily seen by computing the expected value of an observable response for various values of $\theta_j - \delta_i$ using the PF given in Equation 7. Figure 3 shows the expected value of an objective response for the same hypothetical item with four response categories. The categories are coded with the integers 0 to 3 (the codes correspond to the responses of *strongly disagree*, *disagree*, *agree*, and *strongly agree*, respectively). Figure 3 shows that the item elicits greater levels of agreement as the distance between the person and the item on the attitude continuum decreases.

Parameter Estimation

A joint maximum likelihood (JML) approach is used to estimate model parameters. This approach is

Figure 2
Observed Response Category PFs for a Hypothetical Item as a Function of $\theta_j - \delta_i$



based heavily on the (unconditional) procedures described by Andrich (1978), Masters (1982), and Wright & Masters (1982). Assuming that responses are independent across persons and that the responses from any one person are locally independent, then the likelihood function for the GUM may be written as

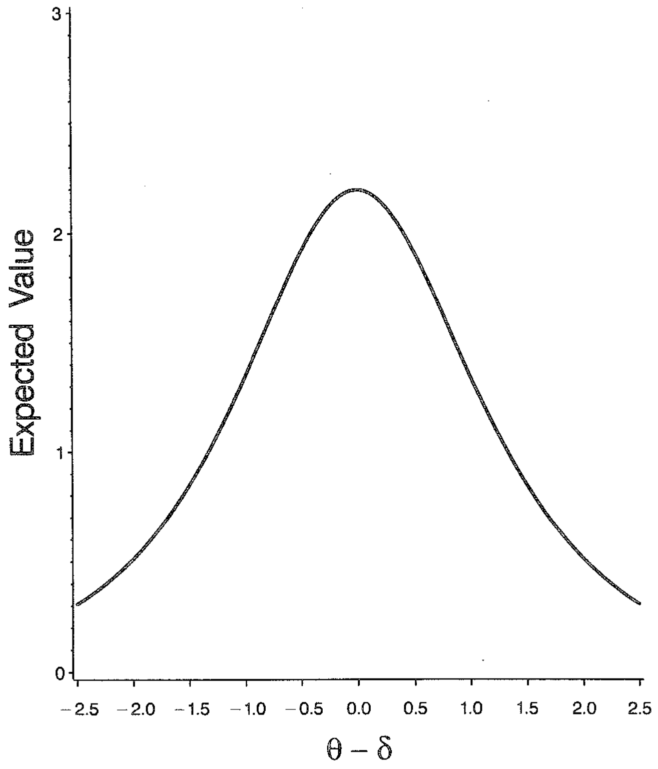
$$L = \prod_{j=1}^{N} \prod_{i=1}^{I} P\left(Z_i = X_{ji} \,|\, \theta_j\right), \tag{9}$$

where X_{ji} is the jth person's observed response to item i. The goal of the maximum likelihood procedure is to find the values of θ_j, δ_i, and τ_k that maximize the probability of obtaining the observed data and, hence, the values that maximize Equation 9. In practice, the natural logarithm of the likelihood equation is maximized instead of the likelihood function itself. The log-likelihood function is equal to

$$\ln(L) = \sum_{j=1}^{N} \sum_{i=1}^{I} \left\{ \left(-\sum_{k=0}^{X_{ji}} \tau_k \right) + \ln\left\{ \exp\left[X_{ji}\left(\theta_j - \delta_i\right)\right] + \exp\left[\left(M - X_{ji}\right)\left(\theta_j - \delta_i\right)\right]\right\} \right.$$
$$\left. - \ln\left\{ \sum_{w=0}^{C} \left[\exp\left(-\sum_{k=0}^{w} \tau_k \right) \right] \left\{ \exp\left[w\left(\theta_j - \delta_i\right)\right] + \exp\left[\left(M - w\right)\left(\theta_j - \delta_i\right)\right]\right\} \right\} \right\}. \tag{10}$$

In the approach presented here, the parameter values that maximize the log-likelihood function are deter-

Figure 3
Expected Values of an Observable Response to a Hypothetical Item as a Function of $\theta_j - \delta_i$



mined in an iterative fashion, following the logic of Wright & Masters (1982). First, the value of τ_1 that maximizes the log-likelihood function is determined, and all other parameters are treated as constants. This maximization process is subsequently repeated for each of the remaining τ_k parameters. The log-likelihood function is then maximized with respect to each of the δ_i parameters in succession, after which the location constraint in Equation 8 is imposed. Finally, the log-likelihood function is maximized with regard to each of the θ_j parameters. Maximizing the log-likelihood function with respect to every parameter constitutes one maximization cycle within the iterative procedure. Maximization cycles are repeatedly performed until the average change in δ_i and τ_k parameters is less than some arbitrarily small value (e.g., .005).

Within any cycle, maximization of the log-likelihood function with respect to a given parameter, Φ, is accomplished using the Newton-Raphson algorithm. The Newton-Raphson algorithm finds the value of Φ at which the partial derivative of the log-likelihood function with respect to Φ is 0. This root is determined iteratively, such that the value of Φ at iteration $q + 1$ is equal to

$$\Phi_{q+1} = \Phi_q - \frac{\left[\dfrac{\partial \ln(L)}{\partial \Phi}\right]_q}{\left[\dfrac{\partial^2 \ln(L)}{\partial \Phi^2}\right]_q}. \qquad (11)$$

Thus, the first- and second-order partial derivatives of the log-likelihood function with respect to Φ are required in order to estimate GUM parameters. These partial derivatives are more easily computed if the log-likelihood function is rewritten as

$$\begin{aligned}\ln(L) = \sum_{j=1}^{N}\sum_{i=1}^{I}\Bigg\{&\left[-\sum_{k=0}^{C} U_{X_{ji}k}(\tau_k)\right] + \ln\left\{\exp\left[X_{ji}(\theta_j-\delta_i)\right] + \exp\left[(M-X_{ji})(\theta_j-\delta_i)\right]\right\}\\ &-\ln\left\{\sum_{w=0}^{C}\left\{\exp\left[-\sum_{k=0}^{C} U_{wk}(\tau_k)\right]\right\}\left\{\exp\left[\mathrm{w}(\theta_j-\delta_i)\right] + \exp\left[(M-w)(\theta_j-\delta_i)\right]\right\}\right\}\Bigg\},\end{aligned} \tag{12}$$

where $U_{X_{ji}k}$ and U_{wk} are dummy variables that are equal to 1 when $k \le X_{ji}$ or $k \le w$, respectively, and are equal to 0 otherwise. The first-order partial derivative with respect to θ_j is then

$$\begin{aligned}\frac{\partial \ln(L)}{\partial \theta_j} &= \sum_{i=1}^{I}\left\{\frac{b_{ji}(X_{ji}) + c_{ji}(M-X_{ji})}{b_{ji} + c_{ji}} - \frac{\sum_{w=0}^{C} d_w\left[e_{wji}w + f_{wji}(M-w)\right]}{\gamma_{ji}}\right\}\\ &= \sum_{i=1}^{I}\left\{\frac{\dfrac{\left[b_{ji}(X_{ji}) + c_{ji}(M-X_{ji})\right]a_{ji}}{\gamma_{ji}}}{P(Z_i = X_{ji})} - \frac{\sum_{w=0}^{C} d_w\left[e_{wji}w + f_{wji}(M-w)\right]}{\gamma_{ji}}\right\}\\ &= \sum_{i=1}^{I}\left\{\frac{P(Y_i = X_{ji}|\theta_j)(X_{ji}) + P(Y_i = M-X_{ji}|\theta_j)(M-X_{ji})}{P(Z_i = X_{ji})}\right.\\ &\qquad\left. - \left[\sum_{w=0}^{C} P(Y_i = w|\theta_j)(w) + P(Y_i = M-w|\theta_j)(M-w)\right]\right\}\\ &= \sum_{i=1}^{I}\left[\mathrm{E}(Y_i|\theta_j, X_{ji}) - \mathrm{E}(Y_i|\theta_j)\right],\end{aligned} \tag{13}$$

where

$$a_{ji} = \exp\left(-\sum_{k=0}^{C} U_{X_{ji}k}\tau_k\right), \tag{14}$$

$$b_{ji} = \exp\left[X_{ji}(\theta_j - \delta_i)\right], \tag{15}$$

$$c_{ji} = \exp\left[(M - X_{ji})(\theta_j - \delta_i)\right], \tag{16}$$

$$d_w = \exp\left(-\sum_{k=0}^{C} U_{wk}\tau_k\right), \tag{17}$$

$$e_{wji} = \exp\left[w(\theta_j - \delta_i)\right], \tag{18}$$

$$f_{wji} = \exp\left[(M-w)(\theta_j - \delta_i)\right], \tag{19}$$

and

$$\gamma_{ji} = \sum_{w=0}^{C} d_w \left(e_{wji} + f_{wji}\right). \tag{20}$$

Note that in Equation 13, $\mathrm{E}\left(Y_i \middle| \theta_j\right)$ is the expectation of the jth person's subjective response to item i, and $\mathrm{E}\left(Y_i \middle| \theta_j, X_{ji}\right)$ is the conditional expectation of the jth person's subjective response given that person's observed response. Similarly, the partial derivatives with respect to δ_i and τ_k are

$$\frac{\partial \ln(L)}{\partial \delta_i} = -\sum_{j=1}^{N}\left[\mathrm{E}\left(Y_i \middle| \theta_j, X_{ji}\right) - \mathrm{E}\left(Y_i \middle| \theta_j\right)\right] \tag{21}$$

and

$$\frac{\partial \ln(L)}{\partial \tau_k} = -\sum_{j=1}^{N}\sum_{i=1}^{I}\left[U_{X_{ji}k} + \frac{\sum_{w=0}^{C} d_w(-U_{wk})\left(e_{wji} + f_{wji}\right)}{\gamma_{ji}}\right] = -\sum_{j=1}^{N}\sum_{i=1}^{I}\left[U_{X_{ji}k} - \sum_{w=0}^{C}(U_{wk})\,P(Z_i = w)\right]. \tag{22}$$

The second-order partial derivatives of the log-likelihood equation are equal to

$$\begin{aligned}\frac{\partial^2 \ln(L)}{\partial \theta_j^2} &= \sum_{i=1}^{I}\left\{\left\{\frac{\left(b_{ji} + c_{ji}\right)\left[b_{ji}X_{ji}^2 + c_{ji}\left(M - X_{ji}\right)^2\right] - \left[b_{ji}X_{ji} + c_{ji}\left(M - X_{ji}\right)\right]^2}{\left(b_{ji} + c_{ji}\right)^2}\right\}\right.\\ &\qquad \left. - \left\{\frac{\left\{\sum_{w=0}^{C} d_w\left[e_{wji}w^2 + f_{wji}\left(M - w\right)^2\right]\right\}\gamma_{ji} - \left\{\sum_{w=0}^{C} d_w\left[e_{wji}w + f_{wji}(M - w)\right]\right\}^2}{\gamma_{ji}^2}\right\}\right\}\\ &= \sum_{i=1}^{I}\left\{\left[\mathrm{E}\left(Y_i^2 \middle| \theta_j, X_{ji}\right) - \mathrm{E}\left(Y_i \middle| \theta_j, X_{ji}\right)^2\right] - \left[\mathrm{E}\left(Y_i^2 \middle| \theta_j\right) - \mathrm{E}\left(Y_i \middle| \theta_j\right)^2\right]\right\}\\ &= \sum_{i=1}^{I}\sigma^2_{Y_i|\theta_j,\, X_{ji}} - \sigma^2_{Y_i|\theta_j}\,, \end{aligned} \tag{23}$$

$$\frac{\partial^2 \ln(L)}{\partial \delta_i^2} = \sum_{j=1}^{N}\sigma^2_{Y_i|\theta_j,\, X_{ji}} - \sigma^2_{Y_i|\theta_j}\,, \tag{24}$$

and

$$\frac{\partial^2 \ln(L)}{\partial \tau_k^2} = \sum_{j=1}^{N}\sum_{i=1}^{I}\left\{\left[\sum_{w=0}^{C} U_{wk}P(Z_i = w)\right]^2 - \left[\sum_{w=0}^{C} U_{wk}P(Z_i = w)\right]\right\}, \tag{25}$$

where

$\mathrm{E}\left(Y_i^2 \middle| \theta_j\right)$ is the expectation of a person's squared subjective response to item i,

$E\left(Y_i^2 \middle| \theta_j, X_{ji}\right)$ is the conditional expectation of a person's squared subjective response to item i given the person's observable response,

$\sigma^2_{Y_i|\theta_j}$ is the variance of a person's subjective response to item i, and

$\sigma^2_{Y_i|\theta_j, X_{ji}}$ is the conditional variance of a person's subjective response to item i given their observable response to item i.

Local Maxima of the Log-Likelihood Function

The log-likelihood function of the GUM with respect to either persons or items need not be single peaked. However, when the log-likelihood function contains local maxima with respect to δ_i, these maxima are usually located relatively far from the point at which the global maximum is attained. Furthermore, analyses of both simulated and real data have indicated that the values of the log-likelihood function at these local maxima are usually much smaller than the global maximum value. Under these circumstances, the Newton-Raphson algorithm appears to perform adequately provided that judicious initial values for δ_i are used. (A strategy for developing initial estimates is described in the Appendix.)

Experience with the GUM has shown that when local maxima occur in the log-likelihood function with respect to θ_j, they may occur at points on the latent continuum that are relatively close to the location of the global maximum. Moreover, the value of the log-likelihood function at these local maxima can be quite similar to the global maximum value. In these situations, the Newton-Raphson procedure often converges to a local maximum solution. Therefore, the maximization algorithm for person parameters includes a grid search that is conducted at various points in the Newton-Raphson process in an effort to prevent locally optimal solutions (Roberts, 1995).

Extreme Response Patterns

The JML procedure yields infinite attitude estimates for persons who consistently use the most extreme *disagree* category (i.e., for persons who receive a score of 0 on each item). Moreover, this extreme response pattern contains no information about the direction of the person's attitude. For example, a person with this response pattern may possess an attitude that is extremely negative or extremely positive relative to the items under consideration. For these reasons, persons who exhibit this extreme response pattern must be discarded from the estimation process. Additionally, experience has shown that erratic attitude estimates may arise when a person does not endorse any item to at least a slight extent. Therefore, those persons who fail to use any response categories that reflect agreement are also discarded from the estimation procedure.

The JML procedure will also produce infinite location estimates for items that consistently elicit the most extreme level of disagreement from every person. This situation can occur whenever a questionnaire contains items that are far more extreme than the attitudes represented in a given sample. If this happens, the items in question should be deleted from the estimation procedure.

Standard Errors of GUM Parameter Estimates

Approximate standard errors (SEs) of the GUM parameters can be derived from the second-order partial derivatives of the log-likelihood function (Wright & Masters, 1982). The approximate SE associated with a specific GUM parameter is a function of the expected value of the corresponding second-order partial derivative

$$\hat{\sigma}_\Phi = \left\{-E\left[\frac{\partial^2 \ln(L)}{\partial \Phi^2}\right]\right\}^{-1/2}, \tag{26}$$

where Φ represents the particular parameter in question. This leads to the following approximations:

$$\hat{\sigma}_{\theta_j} = \left\{ -\sum_{i=1}^{I} \sum_{z=0}^{C} P(Z_i = z) \left\{ \left\{ \frac{\left(\tilde{b}_{ji} + \tilde{c}_{ji}\right)\left[\tilde{b}_{ji} z^2 + \tilde{c}_{ji}(M-z)^2\right] - \left[\tilde{b}_{ji} z + \tilde{c}_{ji}(M-z)\right]^2}{\left(\tilde{b}_{ji} + \tilde{c}_{ji}\right)^2} \right\} \right. \right.$$

$$\left. \left. - \left\{ \frac{\left\{ \sum_{w=0}^{C} d_w \left[e_{wji} w^2 + f_{wji}(M-w)^2 \right] \right\} \gamma_{ji} - \left\{ \sum_{w=0}^{C} d_w \left[e_{wji} w + f_{wji}(M-w) \right] \right\}^2}{\gamma_{ji}^2} \right\} \right\} \right\}^{-1/2} \tag{27}$$

$$= \left\{ -\sum_{i=1}^{I} \left\{ \left\{ \sum_{z=0}^{C} \left[P(Z_i = z) \sigma^2_{Y_i|\theta_j, z} \right] \right\} - \sigma^2_{Y_i|\theta_j} \right\} \right\}^{-1/2},$$

$$\hat{\sigma}_{\delta_i} = \left\{ -\sum_{j=1}^{N} \left\{ \left\{ \sum_{z=0}^{C} \left[P(Z_i = z) \sigma^2_{Y_i|\theta_j, z} \right] \right\} - \sigma^2_{Y_i|\theta_j} \right\} \right\}^{-1/2}, \tag{28}$$

and

$$\hat{\sigma}_{\tau_k} = \left[-\frac{\partial^2 \ln(L)}{\partial \tau_k^2} \right]^{-1/2}, \tag{29}$$

where

$$\tilde{b}_{ji} = \exp\left[z\left(\theta_j - \delta_i\right) \right], \tag{30}$$

$$\tilde{c}_{ji} = \exp\left[(M-z)\left(\theta_j - \delta_i\right) \right], \tag{31}$$

and

$\sigma^2_{Y_i|\theta_j, z}$ is the conditional variance of a person's subjective response given that $Z_i = z$.

Note that $\partial^2 \ln(L)/\partial\theta_j^2$ and $\partial^2 \ln(L)/\partial\delta_i^2$ both involve observed data; thus, the expected values of these terms are obtained numerically (i.e., calculated over $z = 0, 1, \ldots, C$) in Equations 27 and 28. However, $\partial^2 \ln(L)/\partial\tau_k^2$ does not contain observed data, so a numerical expectation is not required in Equation 29.

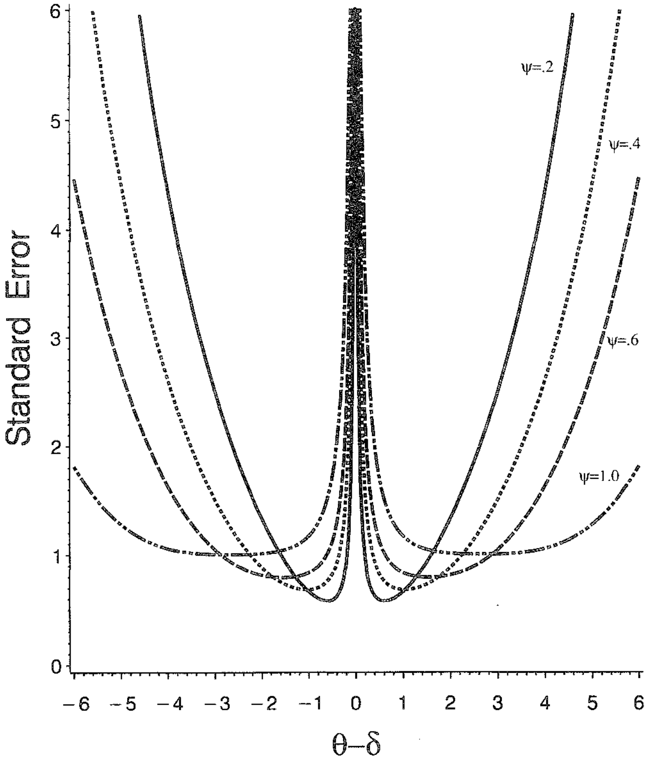
SEs of θ_j estimates ($\hat{\theta}_j$) based on single responses to a single item are plotted in Figure 4. [This is also the SE of a δ_i estimate ($\hat{\delta}_i$) because $\hat{\sigma}_{\theta_j} = \hat{\sigma}_{\delta_i}$ in the case of a single person's response to a single item.] The plot is based on a hypothetical item with six response categories under the assumption that successive τ_k values are equally distant from each other. Each function in Figure 4 corresponds to an alternative value of Ψ, where Ψ is the distance between successive (equally spaced) subjective response category thresholds. The function is symmetric about the origin, and it becomes infinitely large whenever $\left|\theta_j - \delta_i\right|$ is equal to 0 or is infinitely large itself. The global minimum of the function increases as Ψ increases, and the points (or intervals) on the $\theta_j - \delta_i$ axis at which this minimum occurs grow more distant from the origin. Moreover, the range of $\left|\theta_j - \delta_i\right|$ values that yields SEs near the global minimum gets larger as Ψ increases. Thus, SEs become larger, yet more stable, as Ψ increases.

Recovery of Simulated Parameters

Method

Data. The recovery of GUM parameters using the JML procedure was simulated under 30 conditions. Conditions were derived by factorially combining five levels of sample size (100, 200, 500, 1,000, or

Figure 4
SE of a θ_j Estimate as a Function of $\theta_j - \delta_i$



2,000) with six levels of test length (5, 10, 15, 20, 25, or 30). For the varying sample sizes, responses were generated to 5, 10, 15, 20, 25, or 30 attitude items. The δs were always located at equally distant positions on the latent continuum and ranged from −4.25 to +4.25, regardless of the number of items studied. The θs in each condition were randomly sampled from a normal distribution with $\mu = 0$ and $\sigma = 2.0$. The threshold parameters were equally distant, and the interthreshold distance was fixed at .4.

The characteristics of the true item and person parameters used in this simulation were similar to those used in previous studies of estimation accuracy with unfolding item response theory (IRT) models (Andrich, 1988; Andrich & Luo, 1993). The observable response simulated for each person-item combination was on a six-point scale (*strongly disagree*, *disagree*, *slightly disagree*, *slightly agree*, *agree*, and *strongly agree*). Six response probabilities were computed using Equation 7; these were then used to divide a probability interval (i.e., a closed interval between 0 and 1) into six mutually exclusive and exhaustive segments in which each segment corresponded to a particular observable response category. A random number was then generated from a uniform probability distribution, and the simulated response was that response associated with the probability segment in which the random number fell. After an observable response to each item had been generated for all θs, the data were used to estimate GUM parameters. The process of generating data and subsequently estimating parameters was replicated 30 times in each condition, and the true values of all parameters remained constant across replications.

Measures of estimation accuracy. Four measures of estimation accuracy were used. Each measure had been used previously in at least one study of accuracy in IRT parameter estimation (e.g., Andrich, 1988;

Andrich & Luo, 1993; Hulin, Lissak, & Drasgow, 1982; Kim, Cohen, Baker, Subkoviak, & Leonard, 1994; Seong, 1990; Yen, 1987). The root mean squared error (RMSE) provided an index of the average unsigned discrepancy between a set of true parameters and a corresponding set of estimates. The RMSE was calculated across all the parameters of a given type in any single replication. For example, the RMSE of attitude location estimates from a particular replication was computed as

$$\text{RMSE} = \left[\frac{\sum_{j=1}^{N} \left(\hat{\theta}_j - \theta_j\right)^2}{N} \right]^{\frac{1}{2}}, \tag{32}$$

where

θ_j is the true attitude location for the *j*th person,

$\hat{\theta}_j$ is the estimated attitude location for the *j*th person, and

N is the number of persons in the sample.

Similar quantities were computed for the item location parameters and the threshold parameters in a given replication. The RMSE for $\hat{\theta}_j$s can be algebraically decomposed to show that

$$\text{RMSE} = \left[S_{\hat{\theta}}^2 + S_{\theta}^2 - 2\left(S_{\hat{\theta}\theta}\right) + \left(\bar{X}_{\hat{\theta}} - \bar{X}_{\theta}\right)^2 \right]^{1/2}, \tag{33}$$

where

$S_{\hat{\theta}}^2$ is the sample variance of the *N* $\hat{\theta}$s,

S_{θ}^2 is the sample variance of the *N* θs,

$S_{\hat{\theta}\theta}$ is the sample covariance between $\hat{\theta}$ and θ,

$\bar{X}_{\hat{\theta}}$ is the average of the *N* $\hat{\theta}$s, and

$\bar{X}_{\theta}$ is the average of the θs.

(The terms *sample variance* and *sample covariance* are used to designate the biased forms of these statistics in which *N* is used in the denominator rather than *N* – 1.) The RMSE for the $\hat{\delta}_i$s and τ_k estimates ($\hat{\tau}_k$) can be decomposed in an analogous fashion.

Equation 33 shows that the RMSE is sensitive to three types of discrepancies between the true and estimated parameter distributions. First, it depends on the degree of covariation between true and estimated parameter distributions. Relatively large RMSE values are expected when the linear relationship between estimated and true parameters is weak. Second, the RMSE depends on the degree to which the variance of the estimated parameter distribution matches that of the true distribution. It will increase as the variances of the two distributions become more discrepant. Lastly, the RMSE depends on the extent to which the mean of the estimated parameter distribution matches that for the true parameter distribution. It will increase as the absolute difference between the two means grows larger.

The remaining accuracy measures were used to individually evaluate the three discrepancies assessed by the RMSE. A Pearson product-moment correlation between estimated and true parameters was computed on each replication in order to index the degree of covariation between distributions. Similarly, the ratio of estimated parameter variance to true parameter variance was calculated to determine how well the variances of the two distributions matched. Lastly, the mean difference between estimated and true parameters was computed, and the absolute value of this mean difference was used as an estimate of the location difference between the two distributions.

Results

Accuracy of $\hat{\theta}_j$s. Figures 5a–5d show mean accuracy measures associated with the $\hat{\theta}_j$s as a function of the number of persons and number of items simulated in a given condition. All means were based on the 30 replications within a condition. The η^2 values were calculated from a 6 × 5 between-subjects analysis of variance (ANOVA) in which the given accuracy measure was the dependent variable and the number of items and the number of persons were the independent variables. These values indicate the proportion of the corrected total sum of squares that was attributed to the main effect of items (η^2_{I}), the main effect of persons (η^2_{P}), or the interaction of items and persons ($\eta^2_{\mathrm{I \times P}}$).

The RMSEs in Figure 5a and their associated pattern of η^2 values indicated that mean differences in RMSE were almost totally determined by the number of items used to derive the $\hat{\theta}_j$s. The RMSE decreased dramatically as the number of items increased from 5 to 10, and it decreased again when the number of items increased to 15, albeit less substantially. There was little decrease in the RMSE when the number of items increased beyond 20, at which point the average RMSE was .313. This RMSE value was 16% of the average standard deviation (SD) of true θ_j parameters; thus, the amount of error in the $\hat{\theta}_j$s was relatively small compared to the degree of variability in true attitudes.

Average correlations between θ and $\hat{\theta}$ were greater than .959 in every condition (Figure 5b); thus, the $\hat{\theta}_j$ values were practically a linear function of θ. Additionally, the small amount of variability that emerged in these average correlations was almost entirely a function of the number of items used to estimate $\hat{\theta}$. The size of the average correlation grew slightly as the number of items was increased to 20, at which point the average correlation was .992. Further increases in the number of items had little impact on the average correlation between θ and $\hat{\theta}$.

Figure 5c shows the average absolute mean difference between θ and $\hat{\theta}$. This measure was influenced by the number of items, the number of persons, and the interaction of these two factors. However, when 15 or more items were used to calculate $\hat{\theta}$, the average absolute mean difference between θ and $\hat{\theta}$ was always less than .03, regardless of the number of persons. This difference was generally insensitive to further increases in the number of items used.

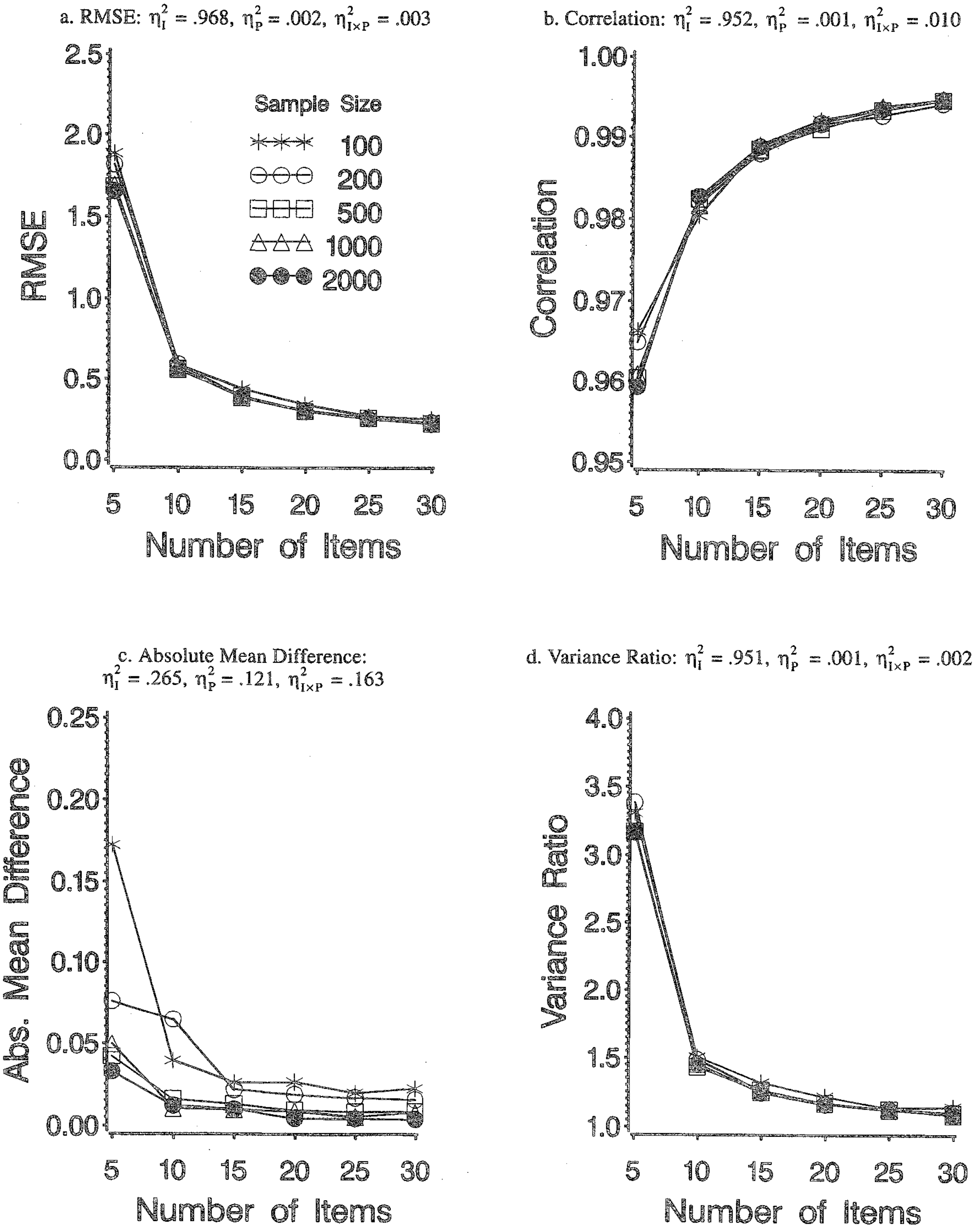
Figure 5d shows the average variance ratio obtained in each condition. The fact that these average ratios were always greater than 1.0 indicated that the variance of $\hat{\theta}$ was consistently larger than the variance of θ. Differences in the variance ratio were almost entirely a function of the number of items ($\eta^2_{\mathrm{I}} = .951$). The variance ratio decreased as the number of items increased to 20, after which further increases in the number of items led to only slight changes. The average variance ratio observed when 20 items were used to estimate θ was 1.178.

Accuracy of $\hat{\delta}_i$s. The accuracy measures for $\hat{\delta}_i$ are shown in Figures 6a–6c. The RMSE and variance ratio measures were both primarily determined by the number of items; these measures showed only small decreases when the number of items increased beyond 20. The average value of the RMSE (Figure 6a) was .217 when 20 items were used; this value was 8.2% of the SD of the δs. The average correlation between true and estimated δ parameters (Figure 6b) was consistently near 1.0 in every condition studied. The average variance ratio was 1.164 when 20 items were used (Figure 6c). For the δ_is, no absolute mean difference measure was used because item locations of both δ and $\hat{\delta}$ were centered at 0.

Although the number of persons had negligible impact on the accuracy of the $\hat{\delta}$ (relative to the impact of the number of items), the effects of persons were in the expected direction. Specifically, the average RMSE and variance ratio generally decreased as N increased.

Accuracy of $\hat{\tau}_k$s. Figures 7a–7d show the accuracy measures for the $\hat{\tau}_k$s derived from the GUM. The η^2s indicate that the RMSEs, absolute mean differences, and variance ratios were all affected by the number of items studied. The values of these accuracy measures consistently decreased as the number of items in-

Figure 5
Mean Accuracy Measures for $\hat{\theta}_j$ s for 5 to 30 Items



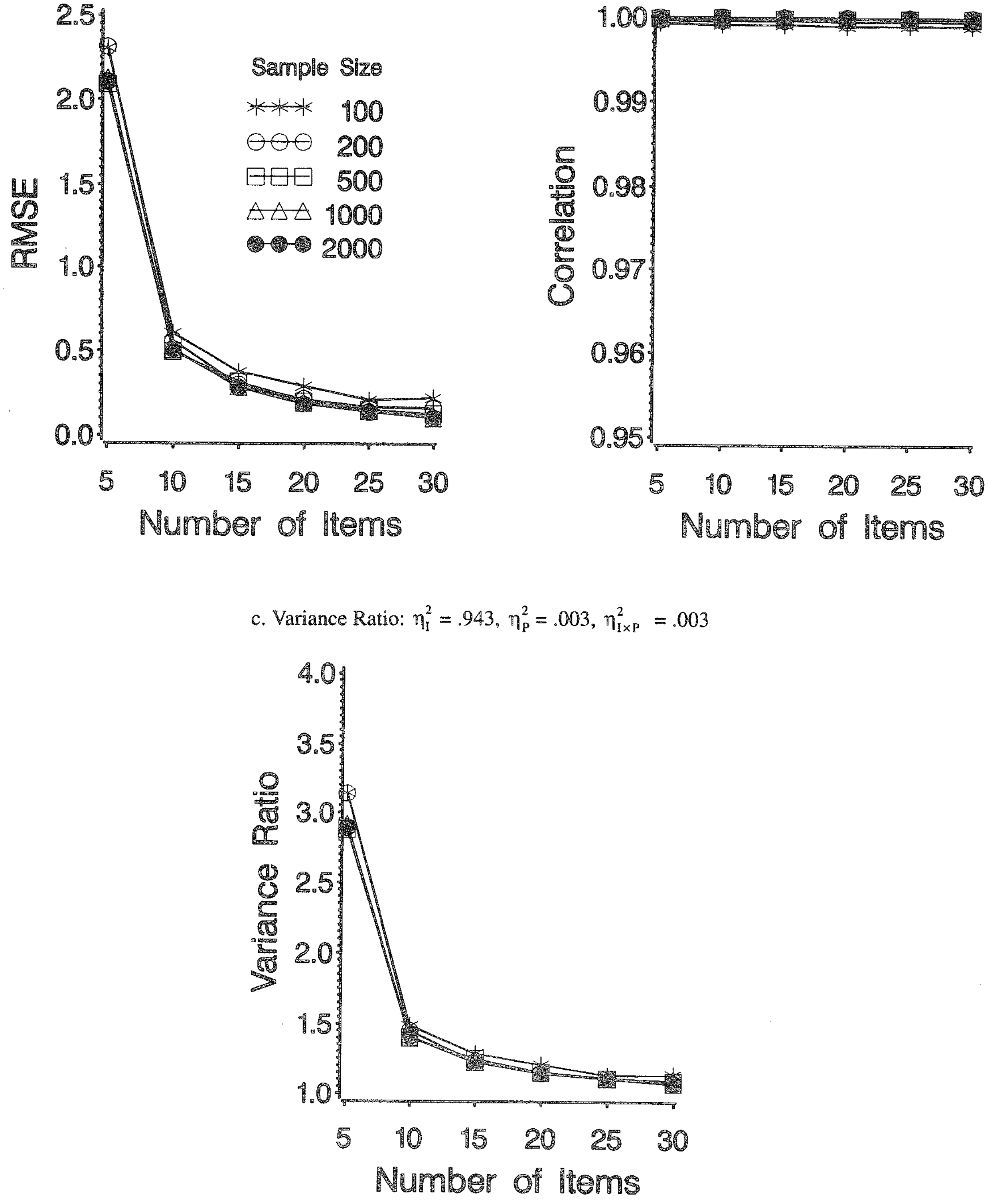
a. RMSE: $\eta_I^2 = .968$, $\eta_P^2 = .002$, $\eta_{I\times P}^2 = .003$

b. Correlation: $\eta_I^2 = .952$, $\eta_P^2 = .001$, $\eta_{I\times P}^2 = .010$

c. Absolute Mean Difference:
$\eta_I^2 = .265$, $\eta_P^2 = .121$, $\eta_{I\times P}^2 = .163$

d. Variance Ratio: $\eta_I^2 = .951$, $\eta_P^2 = .001$, $\eta_{I\times P}^2 = .002$

Figure 6
Mean Accuracy Measures for $\hat{\delta}_i$ s for 5 to 30 Items

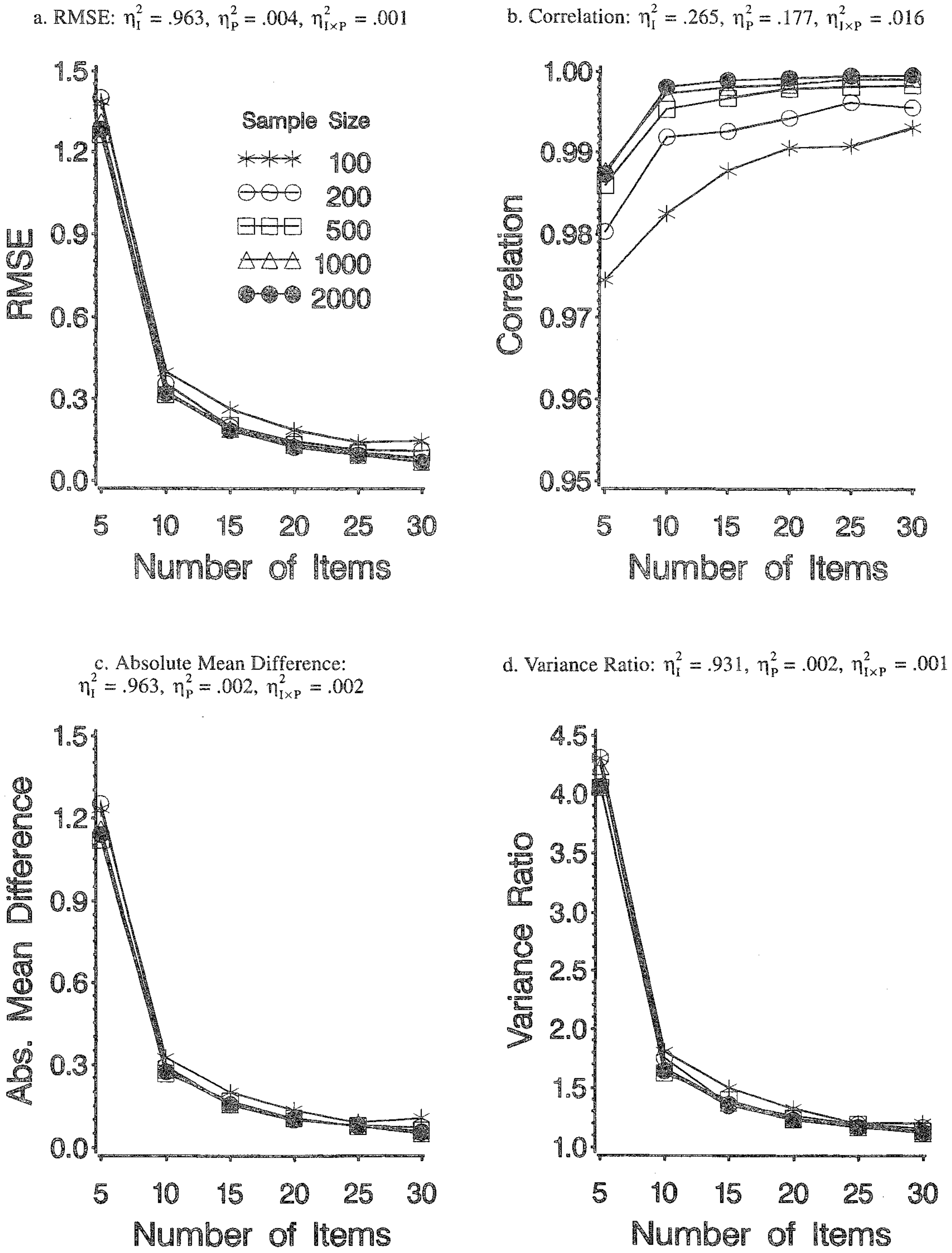


a. RMSE: $\eta^2_I = .959$, $\eta^2_P = .004$, $\eta^2_{I\times P} = .002$

b. Correlation: $\eta^2_I = .005$, $\eta^2_P = .693$, $\eta^2_{I\times P} = .013$

c. Variance Ratio: $\eta^2_I = .943$, $\eta^2_P = .003$, $\eta^2_{I\times P} = .003$

Figure 7
Mean Accuracy Measures for $\hat{\tau}_k$s for 5 to 30 Items



a. RMSE: $\eta_I^2 = .963$, $\eta_P^2 = .004$, $\eta_{I\times P}^2 = .001$

b. Correlation: $\eta_I^2 = .265$, $\eta_P^2 = .177$, $\eta_{I\times P}^2 = .016$

c. Absolute Mean Difference:
$\eta_I^2 = .963$, $\eta_P^2 = .002$, $\eta_{I\times P}^2 = .002$

d. Variance Ratio: $\eta_I^2 = .931$, $\eta_P^2 = .002$, $\eta_{I\times P}^2 = .001$

creased to 20, after which further increases in the number of items had little impact. When 20 items were used, the average RMSE was .143; this value was 22.6% of the SD of true τ_k parameters (Figure 7a). The average absolute mean difference was .111 when 20 items were calibrated (Figure 7c), and the average variance ratio was 1.26 (Figure 7d). Again, the relatively small effect of N was in the expected direction. The three accuracy measures generally improved as N increased.

The average correlation (Figure 7b) between τ_k and $\hat{\tau}_k$ was quite high in every condition (i.e., above .974), but the small amount of variance that did emerge in these average correlations was due to both the number of items studied ($\eta_I^2 = .265$) and N ($\eta_P^2 = .177$). When $N \geq 1{,}000$, the average correlation increased slightly as the number of items approached 10, but further increases in the number of items had little effect on the size of these correlations. With smaller Ns, the correlations continued to increase slightly as the number of items reached 20 to 25.

Discussion

These results indicated that the JML procedure can yield very good linear approximations of GUM parameters. In every condition studied, the correlation between estimated and true parameters was always extremely high. However, the parameter estimates were not always accurate in an absolute sense. When inaccuracies did emerge, they were chiefly due to differences in the variance of the estimated and the true parameters. These inaccuracies were not surprising, given the statistical inconsistency previously seen in JML estimates from other IRT models (Andersen, 1973; Jansen, van den Wollenberg, & Wierda, 1988; Swaminathan & Gifford, 1983; van den Wollenberg, Wierda, & Jansen, 1988; Wright & Douglas, 1977). It is well known that JML estimates of δ_i parameters generally fail to converge to their true values as the number of persons sampled increases. Conversely, JML estimates of θ_j parameters typically fail to converge as the number of test items increases. Although these results did not definitively address the statistical consistency of JML estimates of GUM parameters, they did suggest that any inconsistency will be manifested primarily in the variance of the estimated parameters.

Fortunately, the inaccuracies observed in the JML estimates became quite small when the number of items studied approached 20, and further improvements were generally negligible when test length increased beyond that size. This finding was generally independent of the number of persons used. This does not mean that the number of persons had no effect on parameter estimates. On the contrary, the accuracy of item parameter estimates generally increased as N increased. However, the magnitude of the sample size effect was extremely small relative to the effect of test length, at least within the range of Ns studied here. Therefore, the results of this simulation suggest that accurate estimates of GUM parameters can be obtained in many practical attitude measurement situations in which the attitude questionnaire is based on as few as 15 to 20 six-category items and $N \geq 100$.

An Example With Real Data

Method

Data. Graded responses to Thurstone's (1932) attitude toward capital punishment scale were obtained from 245 University of South Carolina undergraduates. The 24 statements from the scale were presented randomly to each student on a personal computer. Students responded to each statement using one of six response categories—*strongly disagree*, *disagree*, *slightly disagree*, *slightly agree*, *agree*, and *strongly agree*.

Dimensionality. Davison (1977) showed that when graded responses follow a simple metric unfolding model, the principal components of the interitem correlation matrix will suggest two primary dimensions and the component loadings will form a simplex pattern. Simulations of the GUM have suggested that the structure of the interitem correlation matrix will be similar when the model holds perfectly. (The component loadings will form a simplex-like structure, but the endpoints of the simplex will be folded inward.)

Therefore, the interitem correlation matrix was analyzed by a principal components analysis to identify those items that were least likely to conform to the unidimensionality assumption of the GUM. Conformability was operationally defined in terms of the item-level communality estimates derived from the first two principal components. Specifically, an item was discarded if less than 30% of its variation was determined from a two-component solution. This criterion, although somewhat arbitrary, appeared to be reasonable based on previous simulations. Seven items were discarded due to this requirement; this suggested that the original item set was indeed multidimensional. Responses to the remaining 17 items were used to derive initial model estimates.

Eliminating the most poorly fitting items. After an initial calibration of model estimates was obtained, the most poorly fitting items were identified on the basis of two criteria. First, Wright & Masters' (1982) item-fit t statistic (i.e., the infit statistic) was computed for each item. This statistic indexes the degree to which item-level responses fail to conform to the model. Wright and Masters claim that the item-fit t statistic follows a standard t distribution when applied to data that conform to a cumulative model. However, the distribution of the statistic under the GUM is not known. Therefore, this statistic was used simply as a heuristic to identify the most poorly fitting items.

The second criterion used to identify deviant items was based on the degree to which an item's location on the attitude continuum appeared to be inconsistent with its content. Although this type of evaluation was obviously subjective, only those items with clearly questionable locations were removed.

Together, these two heuristics led to the removal of five items. The characteristics of these five items were consistent with the general hypothesis of an ideal point response process, yet their fit to the GUM (i.e., a specific model of that process) was questionable. Responses to the remaining 12 items were used to derive final estimates of model parameters. (Note that in actual test construction situations, a test developer should reconfirm the fit of final scale items using an independent sample of respondents.)

Results

Item location estimates. The item location estimates ($\hat{\delta}_i$) for the 12 selected items are shown in Table 1, along with their approximate SEs. Items are ordered on the basis of their location estimates, and the attitude expressed by each item is consistent with this ordering. The consistency observed between item content and item location provides intuitive support for the adequacy of the model.

There was a pronounced gap between the estimated locations of Items 8 and 9. This was presumably due to a lack of items in the initial pool that reflected intermediate opinions. A scale developer would ordinarily attempt to calibrate a much larger set of items and then select a subset of items from the final calibration that

Table 1
Scale Values ($\hat{\delta}_i$) and Their SEs [$\hat{\sigma}(\delta_i)$] for 12 Capital Punishment Statements From the Final Calibration

	Statement	$\hat{\delta}_i$	$\hat{\sigma}(\delta_i)$
1.	I do not believe in capital punishment under any circumstances.	−1.859	.075
2.	Capital punishment is absolutely never justified.	−1.434	.063
3.	Capital punishment is not necessary in modern civilization.	−1.378	.062
4.	We can't call ourselves civilized as long as we have capital punishment.	−1.298	.060
5.	Execution of criminals is a disgrace to civilized society.	−1.183	.058
6.	Life imprisonment is more effective than capital punishment.	−1.081	.057
7.	I don't believe in capital punishment but I'm not sure it isn't necessary.	−.828	.054
8.	I do not believe in capital punishment but it is not practically advisable to abolish it.	−.604	.053
9.	We must have capital punishment for some crimes.	2.049	.060
10.	Capital punishment is just and necessary.	2.448	.055
11.	Capital punishment gives the criminal what he deserves.	2.463	.055
12.	Capital punishment should be used more often than it is.	2.704	.054

represented alternative regions of the latent continuum in a fairly uniform manner. Although this strategy seems straightforward, some researchers have argued that it is difficult to develop neutral items that are both unambiguous and relevant to the attitude under study (Edwards, 1946; Nunnally, 1978). This argument implies that it will be difficult to develop an item set that uniformly spans the attitude continuum. The generality of this argument is an empirical question that remains to be answered, but recent empirical findings indicate that such problems need not occur (Roberts et al., 1996).

Attitude estimates. The mean $\hat{\theta}$ was 1.229 with a SD of 1.204. The median $\hat{\theta}$ of 1.286 fell between the statements *I do not believe in capital punishment, but it is not practically advisable to abolish it* (Item 8) and *We must have capital punishment for some crimes* (Item 9). Thus, the "average" person in this group begrudgingly supported capital punishment to at least some extent.

Subjective response category threshold estimates. The estimated subjective response category thresholds were $\tau_1 = -2.429$, $\tau_2 = -2.149$, $\tau_3 = -1.783$, $\tau_4 = -1.188$, and $\tau_5 = -.894$. These thresholds were successively ordered and formed a series of intervals on the latent continuum in which a particular subjective response was most likely. The intervals associated with a *strongly agree* response (from either below or above) were relatively wider than those for the remaining subjective responses. However, this result may have been due to the fact that a majority of respondents were located within the gap on the latent continuum bordered by Items 8 and 9; consequently, there was less information about the exact width of the *strongly agree* intervals.

Global model fit. Although model-fit statistics have not yet been developed for the GUM, a descriptive analysis was conducted to gain an intuitive understanding of the global model fit to the capital punishment data. In this analysis, the difference between each person's attitude estimate and each estimated item location (i.e., $\hat{\theta}_j - \hat{\delta}_i$) was calculated. These differences were then sorted and divided into 70 homogeneous groups with 42 differences per group. The GUM expected values and the observed scores (i.e., observable responses) were both averaged across the person-item pairs within each homogeneous group. In this way, much of the random variation was averaged out of the observed scores. A strong linear relationship emerged between average observed scores and the average expected values, as evidenced by a squared correlation of .987. Thus, the GUM fit these data reasonably well.

Figure 8 shows the average expected values and average observed scores as a function of the mean $\hat{\theta}_j - \hat{\delta}_i$ value within each homogeneous group. Figure 8 shows that the GUM arranged both persons and items on the attitude continuum so that the average expected value and the average observed score both increased as the mean difference between person and item locations approached 0. Furthermore, the fit of the GUM appeared reasonable throughout the range of attitudes.

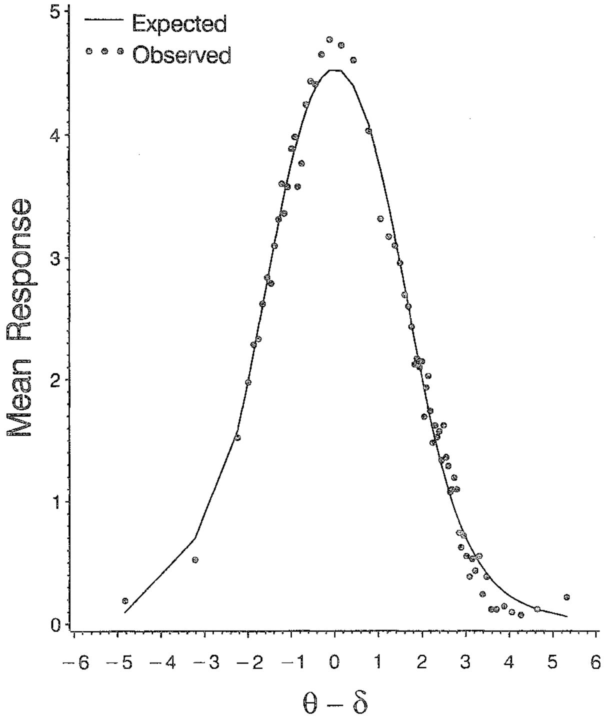
Extensions and Future Research

Alternative Parameterizations of the GUM

Alternative versions of the GUM can be developed simply by reparameterizing the cumulative model that forms the basis for the unfolding process. For example, if it is presumed that subjective responses follow a partial credit model (Masters, 1982) rather than a rating scale model, then a partial credit GUM (PCGUM) can be derived as

$$P\left(Z_i = z \middle| \theta_j\right) = \frac{\exp\left[z\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z} \tau_{ki}\right] + \exp\left[(M - z)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z} \tau_{ki}\right]}{\sum_{w=0}^{C}\left\{\exp\left[w\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w} \tau_{ki}\right] + \exp\left[(M - w)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w} \tau_{ki}\right]\right\}}, \qquad (34)$$

Figure 8
Average Observed Item Responses and Average Expected Item Responses as a Function of $\theta_j - \delta_i$



where τ_{ki} are subjective response category thresholds that are allowed to vary across items and the remaining terms are defined as in Equation 7. This model is appropriate if the scale associated with subjective responses varies across items.

An alternative GUM can be developed by constraining threshold parameters from an underlying rating scale model. Specifically, if it is assumed that successive subjective response category thresholds are equally distant from each other, then subjective responses can be modeled with Andrich's (1982) "constant unit" version of the rating scale model. This is called the constant unit GUM (CUGUM)

$$P\left(Z_i = z \middle| \theta_j\right) = \frac{\exp\left[z(M-z)\lambda + z\left(\theta_j - \delta_i\right)\right] + \exp\left[z(M-z)\lambda + (M-z)\left(\theta_j - \delta_i\right)\right]}{\sum_{w=0}^{C}\left\{\exp\left[w(M-w)\lambda + w\left(\theta_j - \delta_i\right)\right] + \exp\left[w(M-w)\lambda + (M-w)\left(\theta_j - \delta_i\right)\right]\right\}}, \tag{35}$$

where λ is a "unit" parameter that is equal to half the distance between successive thresholds and the remaining terms are defined as in Equation 7. At a conceptual level, this model is appropriate if subjective responses to attitude items are on an equal interval scale. A similar assumption is that subjective responses are on an equal interval scale, but the scale unit changes with each item. In this case, the subjective responses could be modeled with Andrich's (1982) "multiple unit" version of the rating scale model. This would yield a multiple unit GUM (MUGUM),

$$P\left(Z_i = z \middle| \theta_j\right) = \frac{\exp\left[z(M-z)\lambda_i + z\left(\theta_j - \delta_i\right)\right] + \exp\left[z(M-z)\lambda_i + (M-z)\left(\theta_j - \delta_i\right)\right]}{\sum_{w=0}^{C}\left\{\exp\left[w(M-w)\lambda_i + w\left(\theta_j - \delta_i\right)\right] + \exp\left[w(M-w)\lambda_i + (M-w)\left(\theta_j - \delta_i\right)\right]\right\}}, \tag{36}$$

where λ_i is a unit parameter that is allowed to vary across items.

Muraki's (1992) generalized rating scale model can also be used to develop a GUM. The resulting model, referred to as the generalized GUM (GGUM), includes a discrimination parameter that varies across items,

$$P\left(Z_i = z \middle| \theta_j\right) = \frac{\exp\left\{\alpha_i\left[z\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z}\tau_k\right]\right\} + \exp\left\{\alpha_i\left[(M - z)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{z}\tau_k\right]\right\}}{\sum_{w=0}^{C}\left\{\exp\left\{\alpha_i\left[w\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_k\right]\right\} + \exp\left\{\alpha_i\left[(M - w)\left(\theta_j - \delta_i\right) - \sum_{k=0}^{w}\tau_k\right]\right\}\right\}}, \tag{37}$$

where α_i is the discrimination parameter for the ith item. This model is interesting because it allows for item-level response functions that vary in their peakedness and dispersion, but maintain a constant set of subjective response category thresholds across items. In this regard, the GGUM is conceptually similar to Thurstone's successive intervals procedure (Safir, 1937) for scaling attitude items in which the SDs of item scale value distributions are allowed to differ, but the response category boundaries remain fixed. However, the two models differ in that the successive intervals procedure posits a cumulative response mechanism whereas the GGUM is based on an unfolding mechanism.

Roberts (1995) investigated the recovery of parameters from the CUGUM and MUGUM using a JML procedure analogous to that described here. His results paralleled those reported here for the GUM. Specifically, he found that reasonably accurate estimates of all model parameters could be obtained with as few as 15 to 20 six-category items and as few as 100 persons. An investigation of parameter recovery in the PCGUM is currently underway, and preliminary results suggest that relatively large samples (e.g., $N \geq 1{,}000$) are required to produce accurate measures of item-level thresholds.

Alternative Methods of Parameter Estimation

As noted above, the JML procedure yielded reasonably accurate estimates of GUM parameters with minimal data demands. However, the method is computationally intensive, and the statistical consistency of the resulting estimates is generally questionable. Consequently, estimation of GUM parameters using a marginal maximum likelihood procedure (Bock & Aitkin, 1981; Bock & Lieberman, 1970; Muraki, 1992) has been investigated. Preliminary results appear promising and suggest that marginal maximum likelihood estimates may be more computationally efficient and more statistically sound than their JML counterparts.

Conclusions

The GUM provides a way to simultaneously locate both items and persons on a unidimensional latent attitude continuum using an unfolding mechanism; thus, it is consistent with Thurstone's (1928, 1931) view of the response process in attitude measurement. Moreover, it can be used in situations in which responses are binary or graded, so there is no need to dichotomize the data and lose potentially useful measurement information. GUM parameters can be estimated in a reasonably accurate manner with relatively small data demands. Thus, the model should be applicable to a variety of practical attitude measurement situations. Lastly, alternative versions of the GUM can be adapted easily from a diverse set of cumulative IRT models, which adds to the versatility of the approach.

Appendix: Developing Initial Parameter Values

Initial estimates of GUM parameters are derived most easily by first dichotomizing the graded response data such that any level of agreement is scored as 1 and all other responses are scored as 0. The resulting binary scores lead to less cumbersome solutions for initial estimates, and these simpler solutions provide adequate input to the algorithm. For binary scores, there is only one τ_k parameter that must be estimated, and this parameter is denoted τ_B. The initial estimate of τ_B is obtained by first setting all $\theta_j = 0$, and then setting the δ_i value of the most frequently endorsed item to 0. Consequently, the estimate of τ_B must be large enough to account for the proportion of endorsements observed for the most frequently endorsed item. Let h denote the most frequently endorsed item. Presuming that item h is located at $\delta_h = 0$ and all $\theta_j = 0$, then the expected proportion of endorsements for item h is

$$P\left(Z_h{=}1\middle|\theta_j\right) = \frac{\exp(-\tau_B) + \exp(-\tau_B)}{\exp(0) + \exp(-\tau_B) + \exp(-\tau_B) + \exp(0)} = \frac{\exp(-\tau_B)}{1 + \exp(-\tau_B)}. \tag{38}$$

Now, let s_h equal the number of endorsements observed for item h. If the observed proportion of endorsements for item h is set equal to the expected proportion, then an algebraic solution for τ_B can be obtained as

$$\frac{\exp(-\tau_B)}{1 + \exp(-\tau_B)} = \frac{s_h}{N} \tag{39}$$

and

$$\tau_B = -\ln\left[\frac{s_h}{N - s_h}\right]. \tag{40}$$

This estimate of τ_B can be used to derive initial values for both δ_i and τ_k. For example, if it is assumed that all $\theta_j = 0$ and that the location of the item with the largest proportion of endorsements is $\delta_h = 0$, then the absolute values of the remaining item locations can be obtained by solving the following equation with respect to δ_i $(i \neq h)$:

$$P\left(Z_i = 1\middle|\theta_j\right) = \frac{s_i}{N} = 0 = \frac{\exp(-\delta_i - \tau_B) + \exp(-2\delta_i - \tau_B)}{\exp(0) + \exp(-\delta_i - \tau_B) + \exp(-2\delta_i - \tau_B) + \exp(-3\delta_i)} - \frac{s_i}{N}, \tag{41}$$

where s_i refers to the number of endorsements observed for item i. Equation 41 does not have a closed form and must be solved numerically for each δ_i. This is easily accomplished using a bisection algorithm (Press, Teukolsky, Vetterling, & Flannery, 1992). The solutions derived from Equation 41 are the absolute values of the corresponding δ_i estimates. The sign of each estimate is obtained from the signs of the pattern loadings from the first principal component of the interitem correlation matrix (calculated from graded responses). Davison (1977) showed that these signs correspond to the direction of the item when responses exhibit an unfolding structure.

The estimate of τ_B can also be used to calculate initial values for τ_k. If it is assumed that the introduction of a graded response scale will simply decrease the interthreshold distance in a linear fashion and that the distances between successive thresholds are equal, then the following initial estimate of τ_k is reasonable

$$\tau_k = (C + 1 - k)\tau_B / r, \tag{42}$$

where r is the number of observable response categories scored as 1 during the dichotomization process.

Extreme estimates of τ_B should be avoided when calculating initial values. Therefore, τ_B can be restricted to a reasonable range of possible values (e.g., –2.0 to –.5) simply by rescaling the observed proportion of endorsements for each item so that Equation 40 yields an acceptable value. If this is done, then the same rescaled proportions should be used in Equation 41.

Finally, an initial estimate of each θ_j parameter can be obtained by considering those items that a given person has endorsed to at least some extent. A weighted average of the initial δ_i estimates associated with such items can be obtained using the graded response scores as the weights, and this average provides a suitable initial estimate of θ_j.

References

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31–44.

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

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

Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement, 12,* 33–51.

Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49,* 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17,* 253–276.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading MA: Addison-Wesley.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

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

Cliff, N., Collins, L. M., Zatkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement, 12,* 83–97.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika, 42,* 523–548.

Donoghue, J. R. (1994). An examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 41,* 295–311.

Edwards, A. (1946). A critique of "neutral" items in attitude scales constructed by the method of equal appearing intervals. *Psychological Review, 53,* 159–169.

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

Hoijtink, H. (1990). A latent trait model for dichotomous choice data. *Psychometrika, 55,* 641–656.

Hoijtink, H. (1991). The measurement of latent traits by proximity items. *Applied Psychological Measurement, 15,* 153–169.

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.

Jansen, P. G. W., van den Wollenberg, A. L., & Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12,* 297–306.

Kim, S., Cohen, A. S., Baker, F. B., Subkoviak, M. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika, 59,* 405–421.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, No. 140,* 5–53.

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

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

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd ed.). Cambridge: Cambridge University Press.

Rasch, G. (1960/80). *Probabilistic models for some intelligence and attainment tests.* (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980), with foreword and afterword by B. D. Wright. Chicago, The University of Chicago Press.

Roberts, J. S. (1995). Item response theory approaches to attitude measurement. (Doctoral dissertation, University of South Carolina, Columbia, 1995). *Dissertation Abstracts International, 56,* 7089B.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1996). *The Thurstone and Likert approaches to attitude measurement: Two alternative perspectives of the item response process.* Manuscript in preparation, Research Report Series. Princeton NJ: Educational Testing Service.

Safir, M. A. (1937). A comparative study of scales constructed by three psychophysical methods. *Psychometrika, 2,* 179–198.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*

Monograph Supplement, No. 17.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14,* 299–311.

Swaminathan, H., & Gifford, J. A. (1983). Estimation of parameters in the three-parameter latent trait model. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 13–30). New York: Academic Press.

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

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51,* 561–577.

Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33,* 529–554.

Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26,* 249–269.

Thurstone, L. L. (1932). *Motion pictures and attitudes of children.* Chicago: University of Chicago Press.

van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement, 12,* 307–313.

van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists.* Amsterdam: CT Press.

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement, 18,* 97–110.

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281–294.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52,* 275–291.

Acknowledgments

This report was based partially on research from the first author's doctoral dissertation that was conducted at the University of South Carolina under the direction of the second author. We are grateful to Thomas Cafferty, Huynh Huynh, and Douglas Wedell for their critical input during the dissertation process. We would also like to thank John Donoghue and Robert Mislevy for their helpful comments on earlier drafts of this report. Lastly, we have only recently learned that David Andrich has independently proposed an unfolding model for graded responses that is very similar to the GUM (Andrich, 1996). In contrast to this paper, Andrich's development is more conceptual and focuses on using an unfolding model to reconcile the perceived differences between the Thurstone and Likert approaches to measuring social variables.

Author's Address

Send requests for reprints or further information to James S. Roberts, Medical University of South Carolina, Center for Drug and Alcohol Programs (CDAP), Institute of Psychiatry, 472 North, 171 Ashley Avenue, Charleston SC 29425, U.S.A. Email: james_roberts@smtpgw.musc.edu.