DOCUMENT RESUME

ED 476 920 TM 034 971

AUTHOR	Roberts, James S.
TITLE	An Item Fit Statistic Based on Pseudocounts from the Generalized Graded Unfolding Model: A Preliminary Report.
PUB DATE	2003-04-00
NOTE	36p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 21-25, 2003).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Chi Square; *Goodness of Fit; *Test Items
IDENTIFIERS	Unfolding Technique

ABSTRACT

Stone and colleagues (C. Stone, R. Ankenman, S. Lane, and M. Liu, 1993; C. Stone, R. Mislevy and J. Mazzeo, 1994; C. Stone, 2000) have proposed a fit index that explicitly accounts for the measurement error inherent in an estimated theta value, here called chi squared superscript 2, subscript i*. The elements of this statistic are natural byproducts of the marginal maximum likelihood procedure used to estimate generalized grading unfolding model (GGUM) parameters. The objective of this study was to generalize the proposed fit index to the GGUM situation. The behavior of the statistic under the GGUM was examined using alternative simulation techniques, and the results were used to develop hypothesis testing criteria for item misfit assessment. The number of simulated questionnaire items and the number of hypothetical respondents were both varied using a 3 x 4 factorial design, and the number of questionnaire items was 10, 20, or 30, with 500, 750, 1000, or 2000 simulees. Preliminary results suggest that the chi squared superscript 2, subscript i* family of item fit assessment methods is promising in the context of the GGUM. (Contains 1 table, 2 figures, and 21 references.) (SLD)

An Item Fit Statistic Based on Pseudocounts from the Generalized Graded Unfolding Model: A Preliminary Report

James S. Roberts

University of Maryland

RUNNING HEAD: An Item Fit Statistic

Presented at the annual meeting of the American Educational Research Association,

Chicago, Illinois, April 25, 2003.

Introduction

The generalized graded unfolding model (GGUM; Roberts, Donoghue & Laughlin, 2000) is a unidimensional, polytomous item response theory (IRT) model that implements single-peaked, nonmonotonic item characteristic curves (ICCs). It is a proximity-based model that suggests respondents are more likely to receive higher item scores to the extent that they are close to an item on the latent continuum. Researchers have argued that these differences make unfolding models like the GGUM more appropriate for analyzing responses to "Likert-type" questionnaires where the response scale is framed in terms of graded levels of agreement (Andrich, 1996; Roberts, Laughlin, Wedell, 1999; van Schuur & Kiers, 1994). The model is also suitable for certain types of preference measurement and for measuring individual differences within developmental processes that occur in distinct stages (Noel, 1999).

The GGUM defines the probability that the jth respondent will choose the zth response category when responding to the ith item as:

$$P[Z_i = z \mid \theta_j] = \frac{\exp\left(\alpha_i\,[z\,(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\right) + \exp\left(\alpha_i\,[(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_{ik}]\right)}{\sum_{w=0}^{C}\left[\exp\left(\alpha_i[w\,(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\right) + \exp\left(\alpha_i\,[(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\right)\right]} . \tag{1}$$

for z=0, 1… C; where θ_j is the location of the j^{th} individual on the latent continuum, δ_i is the location of the i^{th} item on the latent continuum, α_i is the discrimination parameter for the i^{th} item, τ_{ik} is the k^{th} subjective response category threshold for the i^{th} item, C is the number of response categories minus 1, and M is equal to 2*C+1. In contrast to traditional IRT models with monotonic ICCs, the GGUM yields ICCs that are nonmonotonic and suggest higher item scores to the extent that the distance between θ_j and δ_i approaches zero.

Insert Figure 1 About Here

Figure 1 illustrates item characteristic curves for three hypothetical items under the GGUM. Each curve is centered and symmetric about the point $\delta_i$on the latent continuum. The height and width of each curve are controlled by the α_i andτ_{ik} parameters. The single peaked nature of each curve reflects a proximity-based response process where higher item scores are expected to the extent that the individual is located near the item on the latent continuum. For example, in the case of attitude measurement using a Likert-type questionnaire, the GGUM predicts higher levels of agreement to the extent that the content of the item matched an individual's own attitude. The location of the item on the latent continuum (δ_i) reflects its content, and the location of the individual on the continuum (θ_n) reflects the individual's attitude. Consequently, the GGUM predicts higher levels of agreement (i.e., higher item scores) to the extent that $|\theta_n - \delta_i|$ approaches zero.

The GGUM offers the same advantages provided by other parametric IRT models. Specifically, it offers interpretations of person parameters that are invariant to the items under study, interpretations of item parameters that are invariant to the persons in the sample, and estimates of measurement precision at the individual level. These advantages open the door to applications like computerized adaptive testing (CAT), item banking, and test equating. CAT applications with the GGUM have been studied and appear quite promising (Roberts, Lin & Laughlin, 2001), and several methods to link

GGUM parameter estimates from separate calibrations involving alternative respondent groups have been successfully implemented (Roberts, 2001a). Such applications are generally feasible to the extent that an IRT model fits a given set of responses. However, methods to assess item/model fit have yet to be studied systematically in the GGUM.

The nonmonotonic ICCs that are characteristic of the GGUM lead to a test characteristic function that may possess multiple peaks. Consequently, if an observed test score is obtained simply by summing responses to questionnaire items, then the expected observed score will generally be associated with more than one location on the latent continuum, and the particular locations in question may be quite discrepant. This is also true when the observed score is obtained after reverse scoring responses to negatively worded items presuming that some moderate and/or neutral items are included on the questionnaire. The lack of a one-to-one relationship between true scores and the latent trait makes assessing model/item fit by conditioning on an observed test score more difficult than in the domain of cumulative IRT models. For example, Orlando and Thissen (2000) have developed a method in which expected responses can be quickly aggregated across all response patterns which lead to a given observed test score. These expectations can then be contrasted with observed responses after conditioning on the observed test score. Unfortunately, the effect of nonmonotonicity between the observed score and the latent trait makes this method less attractive, because conditioning on the observed test score does not yield responses associated with individuals who are homogenous with regard to the latent trait.

An alternative to conditioning on the observed score would be to condition instead on the estimated θ value (i.e., $\hat{\theta}$), and compare observed and expected item responses for individuals who are homogenous with respect to $\hat{\theta}$. Chi-squared methods such as those suggested by Bock (1972), Yen (1981), Andrich (1978) typify this approach. However, these methods ignore the fact that $\hat{\theta}$ contains measurement error. As a result, respondents are misclassified into supposedly homogeneous groups to the extent that $\hat{\theta}$ is imprecise. This form of misclassification generally results in Type I errors that may be severely inflated when the number of test items is relatively small (Orlando & Thissen, 2000).

Stone and colleagues (Stone, Ankenmann, Lane & Liu, 1993; Stone, Mislevy & Mazzeo, 1994; Stone, 2000) have proposed a fit index that explicitly accounts for the measurement error inherent in $\hat{\theta}$. This statistic will be denoted here as χ_i^{2*}. The elements of χ_i^{2*} are natural byproducts of the marginal maximum likelihood (MML) procedure used to estimate GGUM parameters, and thus, it is economical to compute. However, the distribution of the index has not yet been analytically derived in the case where polytomous item parameters are estimated from the data (Donoghue & Hombo, 2001). On the other hand, many simulation studies suggest that the statistic is approximated reasonably well by a rescaled chi-squared distribution when data follow from cumulative models (Donoghue, 1998; Hombo & Donoghue, 1999; Stone et al., 1993; 1994; Stone, 2000).

The objective of this paper is to generalize the χ_i^{2*} statistic to the GGUM situation. The behavior of the χ_i^{2*} under the GGUM will be examined using alternative simulation techniques. The results from these simulations will then be used to develop hypothesis testing criteria for item misfit assessment.

The Index of Item Fit Under the GGUM

GGUM item parameter estimates can be easily estimated with a marginal maximum likelihood (MML) approach, and then expected *a posteriori* (EAP) estimates of person locations can be readily obtained. Both of these methods augment the traditional likelihood function by incorporating a continuous prior distribution for θ. In practice, the prior distribution is approximated at a discrete number of points on the latent continuum referred to as quadrature points. The MML estimation procedure yields the following quantities at each of F quadrature points:

$$\bar{r}_{izf} = \sum_{j=1}^{J} \frac{H_{jiz} r_j L_j(V_f) A(V_f)}{\tilde{P}_j} . \tag{2}$$

$$\bar{N}_{if} = \sum_{z=0}^{C} \bar{r}_{izf} . \tag{3}$$

where H_{jiz} is a dummy variable that is equal to 1 when the response to the *ith* item by the *jth* subject is equal to z . Otherwise, H_{jiz} is equal to 0. In Equation 2, V_f is a quadrature point (Stroud & Secrest, 1966), and $A(V_f)$ is the rescaled density of a prior distribution for θ at V_f. The scale of the $A(V_f)$ values is such that:

$$\sum_{f=1}^{F} A(V_f) = 1 . \tag{4}$$

Additionally, $L_j(V_f)$ is the conditional probability of response vector $\mathbf{X}_j$ at quadrature point V_f under the GGUM, and $\tilde{P}_j$ is the marginal probability of response pattern $\mathbf{X}_j$. The quantity $\bar{r}_{izf}$ can be interpreted as the expected number of respondents at quadrature point V_f who receive a score of z on the *ith* item. Similarly, the quantity $\bar{N}_{if}$ can be thought of as the expected number of individuals at quadrature point V_f who respond to the *ith* item. These expectations can be used to derive the χ_i^{2*} statistic as follows:

$$\chi_i^{2*} = \sum_{f=1}^{F} \sum_{z=0}^{C} \left(\frac{(\bar{r}_{ifz} - E_{ifz})^2}{E_{ifz}} \right) \tag{5}$$

where:

$$E_{ifz} = P[Z_i = z \,|V_f]\, \bar{N}_{if} \tag{6}$$

is the expected frequency of response *z* at quadrature point V_f given by the GGUM. Relative to the fit indices proposed by Bock (1972), Yen (1981), and Andrich (1978), the χ_i^{2*} index treats each respondent's location on the latent continuum as a random variable rather than a fixed quantity, and thus, it properly reflects the uncertainty associated with each $\hat{\theta}$.

Unfortunately, the sampling distribution of χ_i^{2*} under the GGUM is unknown. Donoghue and Hombo (1999; 2001a) have determined that χ_i^{2*} is distributed as a quadratic form of normal variables when item parameters are known and either binary and polytomous cumulative models are studied. However, there has been no research to suggest whether this result will hold for unfolding IRT models. Additionally, item parameters are generally unknown, and their estimation complicates the analytical determination of the χ_i^{2*} distribution (Donoghue & Hombo, 2001b).

Given that the theoretical distribution of χ_i^{2*}under the GGUM is unknown, a series of simulations was performed to a) describe the behavior of χ_i^{2*} under both the null hypothesis of perfect fit and a reasonable alternative, and b) determine whether simulation results can be used to develop a practical hypothesis testing strategy based on a rescaled chi-squared distribution.

Simulation Study

Simulation Design and Response Data Generation

The number of simulated questionnaire items and the number of hypothetical respondents were both varied using a 3 x 4 factorial design. The number of questionnaire items was either *I*=10, 20 or 30. True item parameters were sampled from a set of 47 item parameter estimates published by Roberts, Lin and Laughlin (2001). These authors derived the item parameter estimates from responses to an abortion attitude questionnaire with 6 response categories per item (0=strongly disagree, 1=disagree, 2=slightly disagree, 3=slightly agree, 4=agree, 5=strongly agree). The model in Equation 1 was used to obtain the parameter estimates in their study.

The number of simulees generated was either *N*=500, 750, 1000 or 2000. True values of θ were independently sampled from a N(0,1) distribution on every replication. With the true values for persons and items in hand, item responses were then generated using Equation 1. Following the generation of data for a given replication, the item parameters were replaced into the pool for resampling on subsequent replications.

Estimation of Model Parameters and χ_i^{2*}

GGUM parameter estimates were derived from the simulated item response data using a modified version of the GGUM2000 software (Roberts, 2001). The modifications increased the program execution speed but had no impact on the ultimate solution. MML estimates of item parameters were obtained using a N(0,1) prior distribution for θ along with 30 equally-spaced quadrature points ranging from -4 to +4. The iterative estimation algorithm continued until no item parameter changed by more than .001 from one cycle to the next.

The pseudocounts generated after the final iteration of the MML algorithm were stored and later used to calculate the value of χ_i^{2*} for each item using Equation 5. Due to the fact that the pseudocounts were stored with finite precision, there were instances where E_{fz} values were equal to zero. In those instances, the values of E_{fz} and $\bar{r}_{ifz}$ were both incremented by 1E-8. As the results of the simulation were compiled, it became evident that small values of E_{fz} led to distributions of χ_i^{2*} that did not follow a rescaled chi-squared distribution. Specifically, the observed values corresponding to the upper percentiles of the χ_i^{2*} distribution were noticeably too large. Therefore, a second value of χ_i^{2*} was calculated in which .1 was added to every E_{fz} and $\bar{r}_{ifz}$ before applying Equation 5. In order to avoid ambiguity, the original χ_i^{2*} will be called the "uncorrected χ_i^{2*}" value whereas this second calculation will be referred to as a "corrected χ_i^{2*}". A total of 30 replications were conducted in each cell of the factorial design. On a given replication, item response data were generated for N simulees to I items, item parameters were estimated, pseudocounts were saved, and the uncorrected and corrected values of χ_i^{2*} were produced.

Developing Alternative Sampling Distributions for χ_i^{2*}: The Null Hypothesis Case

On a given replication in the null distribution case, the data were generated in accordance with Equation 1, and a model with the same form was used to estimate the item parameters and pseudocounts which could ultimately be used to calculate χ_i^{2*}. Consequently, the data were fit perfectly by the model except for measurement error and the associated χ_i^{2*} values reflected this perfect fit. After the item parameters were estimated on a given replication, they were treated as known (i.e., true) item parameters, and their values were used along with Equation 1 to generate 1000 new item response data sets; each data set having N simulees and I items. These data sets were used to develop sampling distributions for two alternative versions of χ_i^{2*}.

The first sampling distribution for χ_i^{2*} was developed under the assumption that GGUM item parameters were not known. For this case, the GGUM item parameters were re-estimated in each of the 1000 data sets produced on a given replication of the simulation. The resulting item parameter estimates were used along with the simulated item responses to develop pseudocounts which, in turn, provided the necessary input to calculate χ_i^{2*}. This strategy yielded 1000 values of χ_i^{2*} in which the original item parameters were treated as random variables, and these 1000 χ_i^{2*} values formed a sampling distribution under the null hypothesis. To identify χ_i^{2*} in the case where item parameters were re-estimated for each of these 1000 data sets, an additional subscript, E, will be added to the notation ($\chi_{i_B^{2*}}$).

As mentioned earlier, the original values obtained for $\chi_{i_B^{2*}}$ were noticeably skewed in a positive direction, and this skewness was attributed to extremely small E_{fz} values in some cells. Therefore a "corrected" value of $\chi_{i_B^{2*}}$ was calculated in which .1 was added to both the observed and expected values in each cell before calculating Equation 5. This proved to reduce the skew substantially. (Note that alternative constants such as .05, .15 and .20 were explored, but .1 was the smallest constant that yielded a good approximation to a rescaled chi-squared distribution.)

An alternative sampling distribution for χ_i^{2*} was developed in which the original item parameter estimates were used along with the 1000 sets of simulated data to produce

1000 χ_i^{2*} values. In other words, the item parameters were treated as known, and they were not re-estimated when analyzing the 1000 data sets. This method of developing a sampling distribution for χ_i^{2*} decreased calculation time enormously because the estimation of GGUM item parameters required a substantial amount of computer time when calculating $\chi_{i_B^{2*}}$. The generation of pseudocounts was quite fast when known item parameters and item responses were input into the program. The χ_i^{2*} values produced with this procedure will be denoted with a new "K" subscript ($\chi_{i_K^{2*}}$) to designate that the item parameters were treated as known quantities when deriving the statistic.

As noted by Stone (2000), the sampling distributions of either $\chi_{i_B^{2*}}$ or $\chi_{i_K^{2*}}$ derived from cumulative IRT models are generally well approximated by a scaled chi-squared distribution:

$$\chi_i^{2*} \sim \gamma * (\chi^2(\upsilon)) \tag{7}$$

where $\chi^2(\upsilon)$ is a standard chi-squared distribution with υ degrees of freedom and γ is a scale factor. The parameters ν and γ can be estimated using the method of moments as outlined by Stone (2000):

$$E(\chi_i^{2*}) = E\left[\gamma * \chi^2(\upsilon)\right] = \gamma * E\left[\chi^2(\upsilon)\right] = \gamma * \upsilon\ , \tag{8}$$

$$Var\ (\chi_i^{2*}) = Var\ (\gamma * \chi^2(\upsilon)) = 2\gamma^2\upsilon\ , \tag{9}$$

$$\gamma = \frac{Var(\chi_i^{2*})}{2*E(\chi_i^{2*})} \tag{10}$$

$$\upsilon = \frac{E(\chi_i^{2*})}{\gamma} \tag{11}$$

The mean and variance of the sampling distribution for corrected $\chi_{i_B}^{2*}$ was substituted into Equations 10 and 11 to obtain the scale factor, γ, and the degrees of freedom, ν. Then, the observed corrected χ_i^{2*} was rescaled by a factor of $1/\gamma$ and the result was compared to a critical value from a standard chi-squared distribution with ν degrees of freedom. In short, the sampling distribution for corrected $\chi_{i_B}^{2*}$ provided a means to use a standard chi-squared distribution to evaluate the statistical significance of the observed, corrected χ_i^{2*}. This in itself was not pragmatically useful because an enormous amount of computing effort was expended to calculate the sampling distribution of corrected $\chi_{i_B}^{2*}$, from which, ν and γ were derived. However, the sampling distribution of corrected $\chi_{i_K}^{2*}$ was relatively easy to compute. Following Stone (2000), this distribution of corrected $\chi_{i_K}^{2*}$ was used to estimate ν and γ. The observed corrected $\chi_{i_K}^{2*}$ was then rescaled by a factor of $1/\gamma$ and compared to a critical value from a standard chi-squared distribution with ν - s degrees of freedom, where s was equal to the number of estimated item parameters for the *ith* item. Because there were 6 response categories, there were 7 item parameters estimated per item using Equation 1 (i.e., δ_i, α_i, and five τ_{ik} parameters were estimated for each item).

The reader should note that when rescaled statistics were calculated, the corrected (rather than uncorrected) version of the fit statistic was rescaled. This is because Q-Q plots generally suggested that the uncorrected statistics did not follow a chi-squared distribution. The top panel of Figure 2 illustrates a rather typical Q-Q plot for an uncorrected $\chi_{i_B}^{2*}$ distribution and a chi-squared distribution with degrees of freedom equal to the mean of the uncorrected $\chi_{i_B}^{2*}$ values. The extreme positive values obtained in the uncorrected $\chi_{i_B}^{2*}$ distribution degraded its relationship with the chi-squared distribution. In contrast, the Q-Q plot for the corresponding corrected $\chi_{i_B}^{2*}$ distribution exhibited fewer extreme values. It suggested that the corrected $\chi_{i_B}^{2*}$ distribution was better approximated by a rescaled chi-squared distribution than was the uncorrected version of the statistic. Similar plots were observed for $\chi_{i_K}^{2*}$.

Insert Figure 2 About Here

To recapitulate, there were 30 replications in each of 12 cells defined by calibration sample size (500, 750, 1000, or 2000 simulees) and questionnaire length (10, 20 or 30 items). For each item on a given replication, an observed χ_i^{2*} value was generated. Alternative sampling distributions for χ_i^{2*} were generated by assuming that the item parameters underlying the observed value of χ_i^{2*} were estimated ($\chi_{i_B}^{2*}$) or known ($\chi_{i_K}^{2*}$). Calculation of observed χ_i^{2*} and the corresponding sampling distributions for $\chi_{i_B}^{2*}$ and $\chi_{i_K}^{2*}$ was repeated after adding .1 to both the observed and expected values in every term in Equation 5. Each of these sampling distributions was used to derive the parameters ν and γ that were subsequently used to rescale observed χ_i^{2*} values so that they followed a standard chi-squared distribution. A critical value from this chi-squared distribution was used to evaluate the statistical significance of the rescaled observed χ_i^{2*}. Evaluations were performed using nominal Type I error levels of α=.05 and α=.01.

Developing Alternative Sampling Distributions for χ_i^{2*}: The Alternative Hypothesis Case

Recall that the data generated in each of the 30 replications for a given cell in the experimental design was produced using the GGUM defined in Equation 1. Alternative GGUMs can be derived by constraining model parameters. If one constrains the GGUM so that 1) item discriminations are equal to 1 for all items and 2) a common set of thresholds is applicable across all items, then the following "rating scale" form of the GGUM is obtained:

$$P[Z_i = z \mid \theta_j] = \frac{\exp\left([z(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_k]\right) + \exp\left([(M-z)(\theta_j - \delta_i) - \sum_{k=0}^{z} \tau_k]\right)}{\sum_{w=0}^{C}\left[\exp\left([w(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_k]\right) + \exp\left([(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_k]\right)\right]} . \tag{12}$$

In the alternative hypothesis condition, the item responses originally generated with Equation 1 were analyzed using the rating scale version of the model. This produced substantial discrepancies between the response data and the model (i.e., the alternative hypothesis of imperfect model fit was true).

The methods described in the preceding section were applied in the alternative hypothesis scenario in order to evaluate the power of each statistical technique. First, the data that were formerly generated with Equation 1 were analyzed with the model in Equation 12. The pseudocounts from this analysis were used to produce an observed value of χ_i^{2*} for each item. The item parameters from the data analysis were treated as true parameters and were used to generate 1000 independent data sets in which the item responses followed the model in Equation 12. In the case of $\chi^{2*}_{i_B}$, the item parameters were re-estimated for each of these 1000 data sets and the sampling distributions for the corrected and uncorrected $\chi^{2*}_{i_B}$ were calculated from the corresponding pseudocounts. In the case of $\chi^{2*}_{i_K}$, the item parameters were fixed at the values used to generate the data, and pseudocounts were then derived from the 1000 data sets. These pseudocounts were used to calculate the sampling distributions for both the corrected and uncorrected versions of $\chi^{2*}_{i_K}$.

The sampling distributions for the corrected $\chi^{2*}_{i_B}$ and $\chi^{2*}_{i_K}$ statistics were used to estimate ν and γ parameters as outlined in the previous section. The only difference here was that, in the case of $\chi^{2*}_{i_K}$, the degrees of freedom for the chi-squared distribution were approximated as $\nu - 1 - \frac{5}{I}$. This was due to the fact that the 5 threshold parameters (τ_k) were constant across the I items whereas the 1 other item parameter (i.e., δ_i) was allowed to vary across items.

Results

Type I Error

The upper panel of Table 1 portrays the proportion of observed uncorrected χ_i^{2*} values in the null hypothesis condition that exceeded either the 95th or 99th percentile of the uncorrected sampling distribution for $\chi^{2*}_{i_B}$ (i.e., empirical Type 1 error rates). The proportions were obtained by collapsing across all items and replications for a given calibration sample size and questionnaire length condition. When the nominal $\alpha = .05$, the proportion of rejections was generally close to this nominal value. Specifically, the average proportion of rejections across all sample size and questionnaire length conditions was equal to .0505 and there was no systematic deviation from this average as a function of either calibration sample size or questionnaire length. Similarly, the average proportion of rejected null hypotheses when nominal the α=.01 was equal to .0113. Again, there was no systematic influence of sample size or questionnaire length.

The rejection rates reported in the upper panel of Table 1 correspond to the situation where 1) item parameters were treated as random variables that are estimated and 2) there was no correction for extremely small expected values. The second panel of Table 1, reports the corresponding proportion of rejected null hypotheses when observed $\chi_{i_B}^{2*}$ are corrected by adding .1 to the observed and expected pseudocounts in each cell. Again, the empirical rejection rates were close to their nominal values and there was no systematic relationship to either calibration sample size or number of questionnaire items. When the nominal α=.05, the average empirical rejection rate equaled .0470, and when the nominal α=.01, the average empirical rejection rate equaled .0083. These values were slightly smaller than their uncorrected counterparts and tended to underestimate the nominal α very slightly. However, the differences were quite small, and thus, the correction for sparse expected values did not induce a pragmatic change in the empirical Type I error rate.

The third panel in Table 1 gives the empirical Type I error rate achieved when an observed, rescaled, corrected $\chi_{i_B}^{2*}$ statistic was compared to a critical value derived from a chi-squared distribution with ν degrees of freedom. As in the previous panels, the empirical rejection rates were close to their nominal values and the variability was not related to sample size or questionnaire length. The average rejection rate when the nominal α=.05 was equal to .0486 and that for the nominal α=.01 was equal to .0100. Thus, the rejection rates obtained by comparing the observed, rescaled, corrected $\chi^{2*}_{i_B}$ to a critical value from a corresponding chi-squared distribution were very similar to those obtained using a critical value from the sampling distribution of corrected $\chi^{2*}_{i_B}$.

The last panel in Table 1 gives the empirical Type I error rate achieved when an observed, rescaled, corrected $\chi^{2*}_{i_K}$ statistic was compared to a critical value derived from a chi-squared distribution with ν - s degrees of freedom. The empirical Type I error rates underestimated the nominal values by slightly more than 60%. The average empirical rejection rate was equal to .0192 when the nominal α=.05, and it was equal to .0036 when the nominal α=.01. Consequently this method, tended to yield conservative test results in which the null hypothesis was rejected less frequently than expected. Again, the variability of the rejection rates was not dependent on the sample size or the questionnaire length.

Power Rates

Table 2 gives the rejection rates for the situation in which the model in Equation 1 was used to generate the item response data, but the constrained model in Equation 12 was used to analyze the data. In this case, there was an unspecified degree of misfit between the model and the data, although one could speculate that the degree of misfit was pronounced given the variability in item parameters used to generate the data (Roberts et al., 2001). The top panel of Table 2 illustrates the observed power rates obtained by using the 95th or 99th percentile of the uncorrected $\chi^{2*}_{i_B}$ distribution as the critical value for the item fit test. The average empirical power rate was equal to .915 and .738 for the α=.05 and α=.01 cases, respectively. In each case, the observed power increased as the calibration sample size increased. In the case where the α=.05, the power was consistently near or above 80%. However, there was a noticeable increase in power as the sample size increased from 500 to 750, after which, differences for observed power rates were mitigated. In the α=.01 case, the observed proportion of rejections increased steadily as the calibration sample size increased. However, the smaller sample sizes (*N*=500 or *N*= 750) produced somewhat mediocre power rates that were noticeably different from larger sample size conditions.

The second panel of Table 2 provides rejection proportions for the case where the 95th or 99th percentile of the corrected $\chi^{2*}_{i_B}$ distribution was used to determine a critical value for an item fit test. In this case, the empirical rejection rate was .967 when α=.05 and .942 when α=.01. Again, the observed power increased with increasing calibration sample size in all conditions. The conspicuously lower power seen with the uncorrected $\chi^{2*}_{i_B}$ statistic when α=.05 and *N*=500 was not present with the corrected statistic. With the corrected statistic, all observed power rates were equal or greater than .883 when α=.05. Additionally, the relatively mediocre power rates obtained with the uncorrected $\chi^{2*}_{i_B}$ statistic when α=.01 were eliminated with the corrected $\chi^{2*}_{i_B}$ statistic. In the corrected $\chi^{2*}_{i_B}$ case, the smallest power rate achieved when α=.01 was .803. When each test result from the uncorrected $\chi^{2*}_{i_B}$ procedure was compared to that for the corrected $\chi^{2*}_{i_B}$ procedure, the proportion of tests with consistent outcomes equaled .942 when α=.05. This corresponded to a Cohen's κ value of .472. For the α=.01 case, the proportion of consistent test outcomes equaled .784 and κ was equal to .260. Whenever the uncorrected $\chi^{2*}_{i_B}$ procedure led to a rejected null hypothesis, the corrected $\chi^{2*}_{i_B}$ procedure yielded the same test result in all but 3 cases. Thus, the addition of .1 to every cell when calculating $\chi_{i_B}^{2*}$ generally improved the power rate obtained when using the percentiles of its sampling distribution to establish a statistical cutoff. Moreover, this improvement in power occurred with little impact on the corresponding Type I error rate as indicated in the preceding section.

The observed power rates calculated from the corrected $\chi_{i_B}^{2*}$ statistic tended to increase slightly as the number of questionnaire items increased when calibration sample sizes were less than 1000. This behavior was not seen with the uncorrected $\chi_{i_B}^{2*}$ statistic. The cause of this result is not currently known. As the number of informative questionnaire items increases, the posterior distribution for a given respondent's $\hat{\theta}$ will generally become less variable, and thus, $\hat{\theta}$ will be more precise. All of the χ_i^{2*} statistics partition information about observed responses across the latent continuum using each respondent's posterior distribution of θ to allocate the information (i.e., the distribution of pseudocounts is dependent on the posterior distribution of each respondent's θ). Perhaps the increased precision afforded by longer questionnaires is responsible for the effect seen with the corrected $\chi_{i_B}^{2*}$ statistics. However, if this is the case, then it is not clear why the effect is lacking for uncorrected $\chi_{i_B}^{2*}$. Obviously, more work must be performed to better understand this effect.

The third panel of Table 2 gives the observed proportion of rejections when the observed, corrected $\chi_{i_B}^{2*}$ was rescaled and compared to a critical value from a chi-squared distribution. The power rates achieved with this strategy were very similar to those found with the corrected $\chi_{i_B}^{2*}$ approach in the preceding panel for every sample size and questionnaire length condition. When α=.05, the average observed proportion of rejections was equal to .966. The corresponding average was .944 when α=.01. When each test result obtained with the rescaled, corrected $\chi_{i_B}^{2*}$ procedure was compared to that for the corrected $\chi_{i_B}^{2*}$ method, there was a 99% level of consistency in both the nominal α=.05 and nominal α=.01 cases. The corresponding κ values equaled .980 and .962, respectively. These results indicate that if the appropriate scale factor and degrees of freedom are known, then a chi-squared distribution can approximate the rescaled sampling distribution of the corrected $\chi_{i_B}^{2*}$ quite well.

The bottom panel of Table 2 lists the observed rejection proportions using the rescaled, corrected $\chi_{i_K}^{2*}$ procedure. As with the previously described methods, the observed power rates associated with this procedure increased with calibration sample size, and there was a tendency for rates to increase with questionnaire length when the sample size was less than 1000. The average power rates obtained with the rescaled, corrected $\chi_{i_K}^{2*}$ procedure were only slightly smaller than those found when GGUM item parameters were estimated. With the rescaled, corrected $\chi_{i_K}^{2*}$, the average proportion of rejections was equal to .956 and .924 in the α=.05 and α=.01 cases, respectively. When the test outcomes from the rescaled, corrected $\chi_{i_K}^{2*}$ method were individually compared with those from the rescaled, corrected $\chi_{i_B}^{2*}$ procedure, the proportion of consistent decisions equaled .992 when α=.05, and .987 when α=.01. The corresponding κ values were .876 and .873. Thus, with regard to observed power, the rescaled, corrected $\chi_{i_K}^{2*}$ performed similarly to its rescaled, corrected $\chi_{i_B}^{2*}$ counterpart.

Discussion

The empirical Type I error and power rates obtained when using the 95th and 99th percentiles of the corrected $\chi_{i_B}^{2*}$ sampling distribution as statistical decision criteria under the GGUM were both quite reasonable. In some respects, these Type I error and power rates provide a gold standard for what might be expected with a χ_i^{2*} oriented approach to item fit. If the distribution of $\chi_{i_B}^{2*}$ under the GGUM is eventually determined in an analytical fashion, then the application of an analytical procedure will not, in all likelihood, lead to much better Type I error or power rates than those seen here with the corrected $\chi_{i_B}^{2*}$ method. Any analytical specification of the sampling distribution for corrected $\chi_{i_B}^{2*}$ will, no doubt, be an approximation to the empirical sampling distribution that was examined in this study. Nonetheless, most psychometric researchers would prefer the mathematical insight and rigor provided by an analytically derived statistical test for item fit.

It appears that sparse expected frequencies may adversely affect the performance of a $\chi_{i_B}^{2*}$ approach when it is applied to the GGUM. This adverse impact is seen in a reduction of power to detect misfitting items. Sparse expected frequencies may also lead to $\chi_{i_B}^{2*}$ sampling distributions that are not approximated well by a scaled chi-squared distribution. This latter point ultimately limits the utility of the $\chi_{i_B}^{2*}$ approach if one desires to use the chi-squared distribution to approximate it. Therefore, the corrected $\chi_{i_B}^{2*}$ statistic is preferred over its uncorrected counterpart. The corrected $\chi_{i_B}^{2*}$ statistic was approximated well by a scaled chi-squared random variable, and it has similar Type I error and noticeably better power as compared to its uncorrected counterpart.

The primary difficulty in using the sampling distribution of the corrected $\chi_{i_B}^{2*}$ to assess item fit is that a large number of replications are required to produce adequate estimates of the 95th and 99th percentiles. This process is extremely time consuming. For example, the sampling distributions of corrected $\chi_{i_B}^{2*}$ for the 30 item condition with 2000 respondents took over 3.5 days to calculate using a 2.4 GHz computer. Therefore, it is unlikely that this approach will be implemented in most practical measurement situations.

An alternative to the sampling distribution method is to use a rescaled chi-squared distribution to approximate the 95th and 99th percentile points of $\chi_{i_B}^{2*}$. However, a sampling distribution of the corrected $\chi_{i_B}^{2*}$ is, itself, required to estimate the scale parameters of the chi-squared distribution that approximates it. It may be possible to adequately estimate the scale parameters with far fewer than 1000 points from the sampling distribution. If so, then the rescaled, corrected $\chi_{i_B}^{2*}$ method may be useful in practice. This possibility is currently being explored.

From the standpoint of computational efficiency and practical application, the rescaled, corrected $\chi_{i_K}^{2*}$ method of testing item fit under the GGUM may be particularly useful. The Type I error rate for this procedure was moderately conservative, and this characteristic is somewhat problematic. However, it exhibited reasonably good empirical power rates when compared to both the corrected $\chi_{i_B}^{2*}$ and rescaled, corrected $\chi_{i_B}^{2*}$ methods. Calculating the sampling distribution of $\chi_{i_K}^{2*}$ is quite fast given that GGUM item parameters are not re-estimated for each element of the sampling distribution. Moreover, it may be possible to estimate the scale parameters of the corresponding chi-squared distribution with far fewer than 1000 elements in the sampling distribution. This would make the method even faster.

As noted in the title, this paper is a preliminary report, and much more work remains to be done in this study. For example, the power of the χ_i^{2*} methods to detect less serious amounts of misfit should be explored more thoroughly. The misfitting model studied here constrains both α_i and τ_{ik} parameters across all items. Models which constrain only one of these two parameters would presumably show smaller degrees of misfit, and thus, would provide more information about the ability of the χ_i^{2*} methods to detect less conspicuous deviations from the null hypothesis of perfect fit. Additionally, an index that quantifies the amount of misfit that occurs when applying a constrained GGUM model would be desirable. Such an index would facilitate more useful interpretations of the power rates reported here. Further effort should also be devoted to finding an "optimal" constant which is added to each cell in the corrected χ_i^{2*} methods. The optimal constant may be a function of sample size and questionnaire length, and this needs to be determined. The minimal size of the sampling distribution required to adequately estimate scale parameters should also be fully explored. The rescaled, corrected $\chi_{i_B^{2*}}$ and $\chi_{i_K^{2*}}$ methods would both be more practical if accurate estimates of γ and ν were available from more sparse sampling distributions than those studied here. Finally, the relationship that emerged between observed power and questionnaire length when small samples were used with the corrected χ_i^{2*} should be explored further. All of these issues will be addressed in the near future.

As with any simulation study, one should avoid overgeneralizing the results. The simulations presented in this paper were based on item characteristics found in real attitude data, and thus, it is hoped that this will improve the generality of these findings. Additionally, both item characteristics and respondent attitudes were resampled on every replication within the simulation. This should also increase the generalizability of the results. Although further replication is still necessary, these preliminary results suggest that the χ_i^{2*} family of item fit assessment methods are promising in the context of the GGUM.

References

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. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2(4)*, 581-594.

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

Donoghue, J. (1998). *An investigation of the sampling distribution of a likelihood ratio χ^2 measure of IRT drif/DIF.* Paper presented at the 1998 annual meeting of the National Council on Measurement in Education, San Diego, CA

Donoghue, J. R., & Hombo, C. M. (1999, June). *Some asymptotic results on the distribution of an IRT measure of item fit.* Paper presented at the 1999 annual meeting of the Psychometric Society, Lawrence, KS.

Donoghue, J. R., & Hombo, C. M. (2001a, April). *The distribution of an item-fit measure of polytomous items.* Paper presented at the 2001 annual meeting of the National Council on Measurement in Education, Seattle, WA.

Donoghue, J. R., & Hombo, C. M. (2001b, June). *The behavior of an IRT measure of item fit in the presence of item parameter estimation.* Paper presented at the 2001 annual meeting of the Psychometric Society, King of Prussia, PA..

Hombo, C. M., & Donoghue, J. R. (1999). *Predicting the degrees of freedom of an item fit statistic.* Paper presented at the 1999 annual meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Roberts, J. S. (2001a, April). *Equating parameters of the generalized graded unfolding model.* Paper presented at the 2001 annual meeting of the American Educational Research Association, Seattle, Washington.

Roberts, J. S. (2001b). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25,* 38.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement,* 24, 3-32.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26,* 192-207.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement,* **59**, 211-233.

Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement, 25,* 177-196.

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

Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Applications to smoking cessation. *Psychological Methods*, **4**, 173-191.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58-75.

Stone, C. A., Ankenmann, R. D., Lane, S., & Liu, M. (1993, April). *Scaling QUASAR's performance assessment.* Paper presented at the 1993 annual meeting of the American Educational Research Association, Atlanta, GA.

Stone, C. A., Mislevy, R. J., & Mazzeo, J. (1994). *Classification error and goodness-of-fit in IRT models.* Paper presented at the 1994 annual meeting of the American Educational Research Association meeting, New Orleans.

Stroud, A. H., & Secrest, D. (1966). *Gaussian quadrature formulas.* Englewood Cliffs, NJ: Prentice Hall.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Author Notes

This material is based upon work supported by the National Science Foundation under Grant No. 0133019. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Further information can be obtained from James S. Roberts, Department of Measurement, Statistics & Evaluation, University of Maryland, 1230F Benjamin Building, College Park, Maryland 20742. E-mail: jr245@umail.umd.edu.

Table 1. Empirical Type I error rates.

Uncorrected χ_E^{2*}						
	α=.05			α=.01		
Sample	Number of Items			Number of Items		
Size	10	20	30	10	20	30
500	.073	.042	.046	.017	.005	.006
750	.057	.053	.047	.013	.013	.011
1000	.040	.053	.056	.017	.013	.013
2000	.040	.057	.042	.007	.013	.008
Corrected χ_E^{2*}						
500	.053	.035	.037	.007	.005	.010
750	.063	.045	.053	.020	.008	.008
1000	.053	.035	.056	.007	.003	.008
2000	.047	.047	.040	.003	.015	.006

Table 1. Empirical Type I error rates (continued).

Sample Size	α=.05 Number of Items 10	20	30	α=.01 Number of Items 10	20	30
Rescaled χ_E^{2*}						
500	.053	.038	.036	.007	.005	.012
750	.067	.047	.056	.020	.008	.009
1000	.057	.037	.054	.013	.005	.009
2000	.053	.047	.038	.010	.015	.007
Rescaled χ_K^{2*}						
500	.017	.015	.021	.003	.003	.009
750	.017	.017	.030	.003	.003	.003
1000	.023	.007	.024	.003	.003	.004
2000	.013	.025	.021	.000	.008	.001

Table 2. Empirical power rates.

Uncorrected χ_E^{2*}						
	α=.05			α=.01		
Sample	Number of Items			Number of Items		
Size	10	20	30	10	20	30
500	.787	.792	.780	.567	.488	.454
750	.917	.922	.916	.697	.668	.671
1000	.953	.962	.957	.807	.803	.792
2000	1.00	.993	.998	.970	.977	.966
Corrected χ_E^{2*}						
500	.883	.925	.944	.803	.872	.882
750	.943	.967	.983	.923	.942	.968
1000	.977.	.990	.988	.957	.980	.978
2000	1.00	.998	1.00	1.00	.997	.999

Table 1. Empirical power rates (continued).

Sample Size	α=.05 Number of Items			α=.01 Number of Items		
	10	20	30	10	20	30
			Rescaled χ_E^{2*}			
500	.873	.925	.948	.813	.875	.892
750	.943	.968	.983	.923	.945	.968
1000	.980	.990	.988	.960	.982	.979
2000	1.00	.998	1.00	1.00	.997	.999
			Rescaled χ_K^{2*}			
500	.840	.910	.938	.750	.857	.876
750	.933	.962	.981	.880	.933	.964
1000	.940	.990	.984	.890	.973	.979
2000	.997	.998	1.00	.990	.997	.999

Figure 1

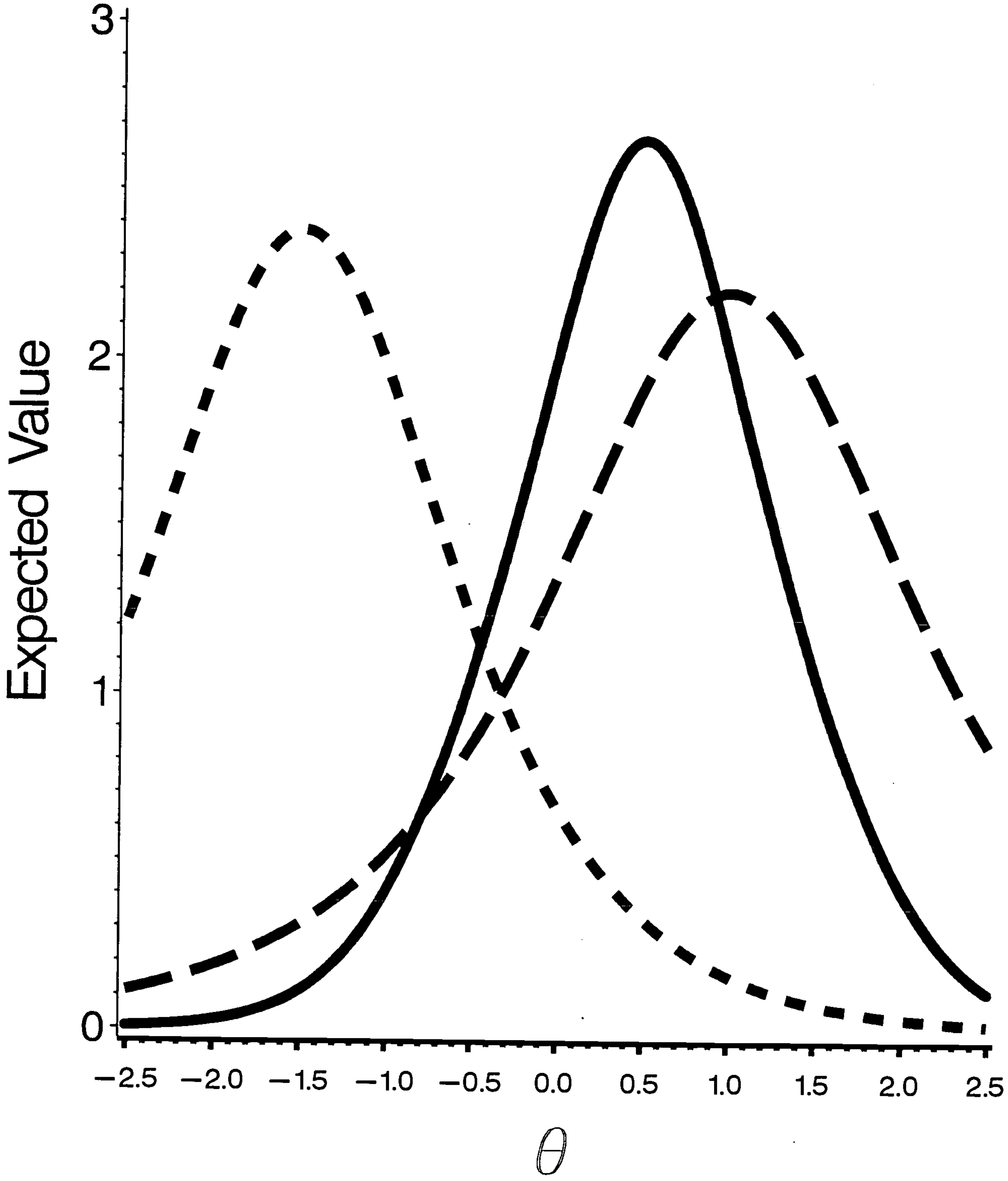
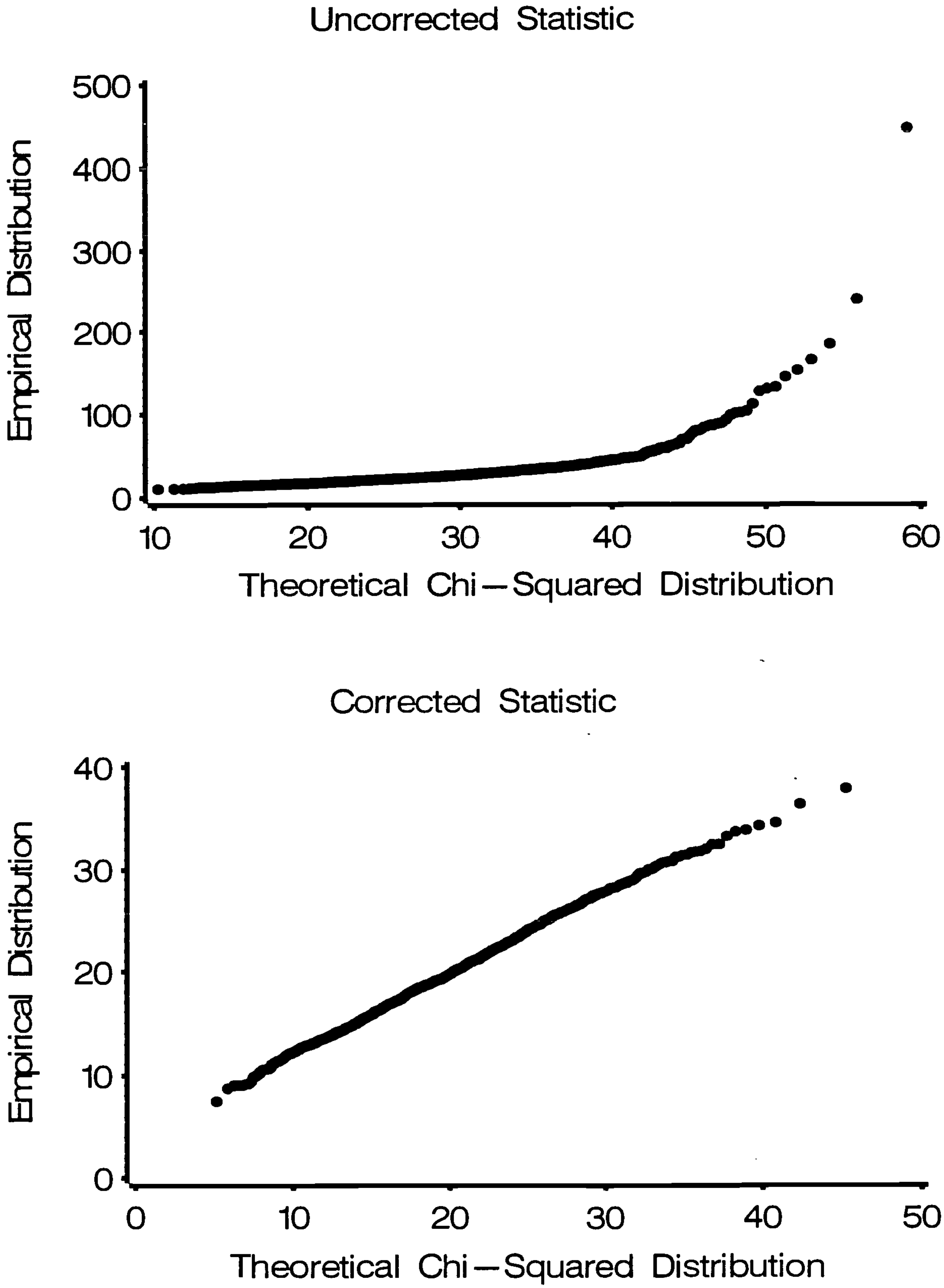


Figure 2



Uncorrected Statistic

Corrected Statistic

Figure Captions

Figure 1. Item characteristic curves for three hypothetical items under the GGUM.

Figure 2. Q-Q plots for an uncorrected χ_E^{2*} (top panel) and a corrected χ_E^{2*} (bottom panel) distribution for a typical item as a function of a theoretical chi-squared distribution.

TM034971

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: An Item Fit Statistic Based on Pseudocounts From the Generalized Graded Unfolding Model: A Preliminary Report

Author(s): James S. Roberts

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ ☑	Level 2A ↑ □	Level 2B ↑ □
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: James S. Roberts

Printed Name/Position/Title: James S. Roberts, Assitant Professor

Organization/Address:
University of Maryland
EDMS
1230 Benjamin Bldg
College Park, MD 20742

Telephone: (301) 405-3630

FAX: (301) 314-9245

E-Mail Address: ↑
JR245@umail.umd.edu

Date: 5/20/03

(Over)