A peer reviewed, open-access electronic journal: ISSN 1531-7714

Assessing Model Fit of the Generalized Graded Unfolding Model

Abdulla Alzarouni, *University of Nebraska-Lincoln*
R. J. De Ayala, *University of Nebraska-Lincoln*

Abstract: The assessment of model fit in latent trait modeling is an integral part of correctly applying the model. Still the assessment of model fit has been less utilized for ideal point models such as the Generalized Graded Unfolding Models (GGUM). The current study assesses the performance of the relative fit indices *AIC* and *BIC*, and the absolute fit adjusted chi-square statistic for the GGUM for both dichotomous and polytomous data. Factors included data generation model, sample size, instrument length, and screening value. Results show that relative fit indices performed well in identifying the GGUM when at least 20-items were used. For polytomous data the correct generation model was identified as the best fitting mode irrespective of the number of items and sample size. The adjusted chi-square statistic performed well in correctly identifying GGUM as the best fit for the GGUM dichotomous data generation, but performed poorly with the dominance models. With polytomous data case these fit indices always correctly identified GGUM as the best fit for the GGUM data. An explanation for this performance is provided.

Keywords: Item Response Theory, Model Fit, GGUM, Attitudes, Measurement

Introduction

Item response theory (IRT) models provide several advantages over classical test theory such as the independence of item and person parameter estimation from calibration samples and the assessment of person parameter estimation accuracy at the individual level. However, these advantages will not be realized if the selected IRT model does not fit the data. Surprisingly, fit determination in applications is not as ubiquitous as one would expect. For example, it has been estimated that more than 40% of published articles in the organizational research literature utilizing IRT models do not include any fit examination (Nye et al., 2020).

There are many IRT models that may be used for proficiency, attitude, or personality assessment. One IRT taxonomy classifies these models as either ideal point or dominance models. The 1-, 2-, 3-parameter models and the graded response model (GRM; Samejima, 1969) are examples of dominance-based models. With dominance models a person's disagree-agree response to an attitude item reflects the extent to which

their opinion (i.e., their location on the latent trait) is greater than the sentiment stated in the item (i.e., item's location) (cf. Roberts & Laughlin, 1996). In contrast, with ideal point models a person's disagree-agree response reflects the extent to which their opinion coincides with the sentiment stated in the item (i.e., person and item locations match). The graded unfolding model and the generalized graded unfolding model (GGUM; Roberts et al., 2000) are examples of ideal point models (Roberts, Donoghue, & Laughlin, 2000). Several researchers (Drasgow et al., 2010; Chernyshenko et al., 2007 (cited in Tay et al., 2011); Nye et al., 2020; Roberts et al., 2000; Stark et al., 2006; Tay et al., 2011) have shown that both dichotomously- and polytomously-scored attitude or personality statements involving self-report are best represented by ideal point models.

Although Roberts (2008) has examined item data fit for ideal point models (GGUM), few studies have systematically examined fit for ideal point models relative to dominance models (Roberts, 2008). A fit statistic commonly used with the GGUM is the adjusted chi-square statistic. Unfortunately, studies examining its performance have shown contradictory results (see Nye et al., 2020; Tay et al., 2011). This study investigates the (absolute fit) adjusted chi-square statistic as well as information criterion (relative) fit approaches (*AIC* and *BIC*) in detecting GGUM misfit with unidimensional dichotomous and ordered polytomous simulated data. Below we briefly introduce the dominance and ideal point models followed by a contrast between the two classes, and a review of the fit literature as it applies to the GGUM.

Literature Review

IRT Dominance Models

One commonly used model for dichotomous unidimensional data is the three-parameter logistic model (3PLM; Birnbaum, 1968). The 3PLM specifies the probability (p) of a response x_i (e.g., correct response/endorsement) on item i given the latent trait of interest (θ) as:

$$p(x_i = 1|\theta, \alpha_i, \delta_i, \gamma_i) = \gamma_i + (1 - \gamma_i) \frac{e^{1.702\alpha_i(\theta-\delta_i)}}{1+e^{1.702\alpha_i(\theta-\delta_i)}}, \quad (1)$$

where α_i is the discrimination parameter, δ_i is an item location parameter, γ_i is the pseudo-guessing parameter; $i = 1...I$ items. By setting $\gamma_i = 0$ we obtain the two-parameter logistic model (2PLM). The $p(x_i = 1|\theta, \alpha_i, \delta_i, \gamma_i)$ as a function of θ is represented by an item response function (IRF).

Although multiple IRT models are applicable to ordered polytomous data, our focus in on the graded response model (GRM; Samejima, 1969). The GRM compares response probabilities in a cumulative fashion. Thus, according to the GRM the probability of obtaining a category score x_i or higher on item i conditional on θ is:

$$P^*_{xi} = \frac{e^{\alpha_i(\theta-\delta_{xi})}}{1+e^{\alpha_i(\theta-\delta_{xi})}}, \quad (2)$$

where P^*_{xi} is the cumulative probability, α_i is the discrimination, δ_{xi} is the category boundary location for category score x_i, and $x_i = \{0, 1, ..., m_i\}$; m_i represents the number of category boundaries. The $\delta_{xi=k}$ represents the boundary between categories k and $k-1$. To obtain the probability of a response in a specific category k (p_k) requires taking the difference between successive P^*_{xi}s. For example, to obtain the probability in a specific category k (i.e., $x_i = k$) we have $p_k = p^*_k - p^*_{k-1}$ with $P^*_0 \equiv 1$ and $P^*_{mi+1} \equiv 0$. This probability as a function of θ is represented by an option response function (ORF).

IRT Ideal Point Models

Ideal point models do not assume the dominance models' cumulative monotonic response function, but rather a non-cumulative unfolding single-peaked response function (Roberts et al., 2000) that reflects that a person's response depends on the proximity between an item's location and the person's standing on the latent trait. The GGUM is:

$$p(x_i = c \mid \theta_j) = \frac{\exp\{\alpha_i [c(\theta_j - \delta_i) - \sum_{k=0}^{c} \tau_{ik}]\} + \exp\{\alpha_i [(M-c)(\theta_j - \delta_i) - \sum_{k=0}^{c} \tau_{ik}]\}}{\sum_{w=0}^{B}\{\exp\{\alpha_i [w(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\} + \exp\{\alpha_i [(M-w)(\theta_j - \delta_i) - \sum_{k=0}^{w} \tau_{ik}]\}\}}, \quad (3)$$

where x_i represents the observed response to item i, τ_{ik} indicates the location of the kth response category threshold on the latent continuum with respect to the ith item location (Roberts et al., 2000), with $c = 0$ and $c = B$ indicating the strongest level of disagreement and agreement, respectively, B is the number of observable response categories minus 1 (i.e., $c = 0, 1, 2, ..., B$), $M = 2B + 1$, and all other symbols are defined above.

Dominance and Ideal Points Models' Item Response Functions

The use of the 2PLM with self-report data such as those assessing attitudes and personality (i.e., noncognitive items) can be found in the literature (e.g., Tay et al., 2011). As mentioned above, these data may also be represented by an ideal point process (Roberts & Laughlin, 1996). In this regard, Tay et al. (2011) found "… the GGUM fits dominance data about as well as the 2PLM *short scales* and is only slightly inferior for long scales" (p. 287; italics ours); "short" scales are defined as 15 items.

Figure 1. Response Functions (Left: GGUM, $\alpha = 0.9$, $\delta = 2$, $\tau_1 = -1.3$; 2PLM, $\alpha = 1.1$, $\delta = 0.5$) and GGUM ORFs (Right: GGUM, $\alpha = 0.7$, $\delta = 0.3$, $\tau_1 = -2$, $\tau_2 = -0.8$, $\tau_3 = -0.3$).

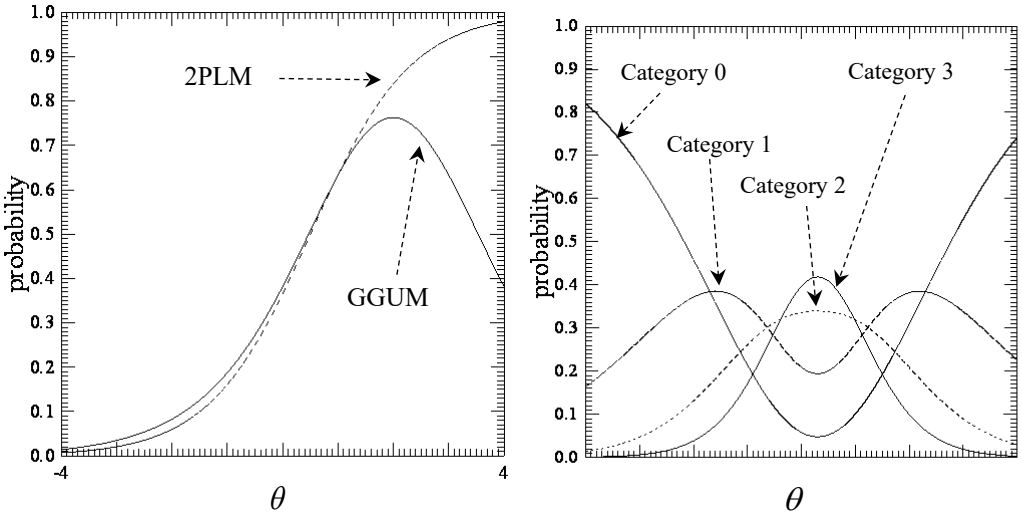


Figure 1 (left panel) shows the GGUM and 2PLM IRFs for a hypothetical dichotomously scored item. For instance, assume we have an item "When can a woman have an abortion?" with responses "Never" and "Anytime." We score "Never" as 0, "Anytime" as 1. Our construct's continuum runs from less/not favorable at the low end to favorable at the upper end. As can be seen, the GGUM IRF shows that the probability of endorsing this item has an ideal location on the continuum (i.e., around 2) and that as one progresses away from this location in either direction the probability decreases. In contrast, the 2PLM IRF predicts that as a person's location increases the probability of endorsing this item increases. Thus, to the extent that the observed data at the upper end on the continuum reflect individuals who believe abortions are permissible only under certain circumstances (e.g., rape/incest), the 2PLM will not appropriately model the data. In

contrast, this data pattern can be correctly modeled with the GGUM. As can be seen, the GGUM and 2PLM response functions provide very similar response probabilities for persons located below approximately 1.5. This similarity of GGUM and 2PLM IRFs is common except for large item locations (e.g., $\delta > 3$). In other words, unless the GGUM location parameters are extreme enough for their response probabilities to be differentiable from those of dominance models, dominance IRT models such as the 2PLM can fit GGUM data well for items with low to moderate item locations (e.g., $\delta < 1.5$; see Figure 1 (left panel)).

Figure 1 (right panel) presents the GGUM's ORFs for a hypothetical ordered polytomously-scored item. To provide context assume that we are interested in measuring the need for cognition, which is "the tendency for an individual to engage in and enjoy thinking" (Cacioppo & Petty, 1982, p. 116). The continuum has individuals who would score high on the need for cognition at one end (e.g., faculty, researchers, physicians, nurses) and individuals who would score low at the other end such as construction workers and "hypothetically" some intellectuals (e.g., philosophers) who might argue that mental deliberation doesn't always lead to the best course of action (Arpaly & Schroeder, 2018). This latter group might in turn disagree to statements that would lead to high scores on the inventory. An example item from the scale is "I tend to set goals that can be accomplished only by expending considerable mental effort" with a four-point response scale of "almost never true," "rarely true," "often true," and "almost always true." As can be seen, the probability of selecting category 0 (e.g., "almost never true") is the highest for individuals at the lower and upper continuum. Similarly, category 1 ("rarely true") is the most likely response for individuals located at approximately -1.5 and 1.5. Respondents located in the center of the continuum are most likely to select category 3 ("almost always true") although category 2 is still a stronger possibility than either category 0 or 1. Conditional on θ the probabilities of responding to each category sum to 1.

Model Fit Statistics

One common fit approach involves the squared residuals between the observed and predicted responses to determine the degree of misfit between the data and fitted model. The process of examining and comparing residuals in IRT for item/model fit examination usually involves chi-square or likelihood-ratio tests (Ames & Penfield, 2015). As mentioned above, research involving these absolute fit indices has, in general, not yielded consistent results in identifying GGUM misfit.

Drasgow et al. (1995) presented a family of absolute fit statistics known as the chi-square statistic. One can calculate these statistics for an item or for multiples of items. The former is referred to as an item single, whereas the latter can involve two, three, or more items and is discussed below. The general form of the chi square fit statistic for item singles and dichotomous data is:

$$\chi_i^2 = \sum_{k=0}^{1} \frac{[O_i(k) - E_i(k)]^2}{E_i(k)}, \quad (4)$$

where $O_i(k)$ is the observed frequency of option k and $E_i(k)$ represents the expected number of respondents selecting option k. $E_i(k)$ is obtained by $E_i(k) = N \int P(x_i = k|\theta) f(\theta) d(\theta)$, where f refers to the θ density (e.g., a unit normal), and the integration uses 161 quadrature points across [-3, 3] (Drasgow et al., 1995). That is, these statistics calculate the expected response frequencies based on an assumption of a unit normal distribution.

An item double is a χ^2 statistic based on the expected frequency of endorsing response options k and k' concurrently (i.e., expected frequency for an item pair in the $(k, k')^{th}$ cell of the two-way contingency table for items i and j, respectively). After determining the observed frequencies for items i and j from a two-way contingency table the expected frequencies are obtained by:

$$E_i(k, k') = N \int P(x_i = k|\theta) P\left(x_j = k'\middle|\theta\right) f(\theta) d(\theta), \quad (5)$$

In a similar fashion Equation 5 can be extended to obtain a χ^2 for item triples. For example, a three-way contingency table is used for estimating the χ^2 using item triplets (see Tay et al., 2011). There are $\binom{I}{2}$ χ^2 possible statistics for item doubles and $\binom{I}{3}$ χ^2 for item triples. Because the possible combinations of item doubles and triples exponentially as the number of items increases, Drasgow et al. (1995) divided the *I* test items into $I/3$ sets of three items. These sets were then used to compute the corresponding χ^2 statistics for individual items, item pairs for doubles, and the whole set for triples. The degrees of freedom (*df*) for these χ^2 statistics equals the number of cells minus one. For instance, for an item with three response categories there are two *df*s and for an item double where each item has 3 response categories the $df = 3*3 - 1 = 8$.

To account for the dependency of χ^2 on sample size as well as to enhance the comparability across different sample sizes the χ^2 for item singles, doubles, and triples are *adjusted* to a sample size of 3,000 (Tay et al., 2011), referred to as the (adjusted χ^2) chi-square statistics. The fit statistic for such items is the ratio of the chi-square to the respective *df*: χ^2/df (i.e., the normed chi-square). Therefore, with the sample size adjustment we have essentially a modified noncentrality parameter estimate:

$$adj\ \chi_i^2 = 3{,}000\frac{\chi_i^2 - df}{N} + df, \tag{6}$$

The $adj\ \chi_i^2$ ratio fit statistic for item singles, doubles, and triples has also been extended to assessing model-level fit by aggregating the item-level statistics. The premise involves taking the mean of the $adj\ \chi_i^2$ ratios and comparing it with the value of 3. Mean ratios less than 3 for item singles, doubles, and triples indicate good model fit (Chernyshenko et al., 2001); this criterion also applies to the χ^2/df ratios (Drasgow et al., 1995). The value of 3 is based on empirical findings using empirical large cognitive ability and personality data sets with dominance models (Chernyshenko et al., 2001; Drasgow et al., 1995). Additionally, their design did not allow an investigation of the statistic's ability to identify known misfit/fit nor was the justification for a value of 3 articulated. Although based on a normed chi-square it should be noted that the $adj\ \chi_i^2$ ratios and mean $adj\ \chi_i^2$ ratios may be negative. Thus, this statistic does not follow a chi-square distribution.

Research has shown that the chi-square statistic ratio for item singles is generally insensitive to detecting misfit under various conditions. For example, Nye et al. (2020) and Tay et al. (2011) found that a chi-square statistic ratio for item singles is a poor indicator of misfit (i.e., predicated on a value of 3) under most conditions pertaining to different sample sizes and number of items. However, the use of doubles and triples have, generally speaking, have had mixed results. For instance, Tay and colleagues (2011) found the $adj\ \chi_i^2$ ratio fit tests for item pairs and triplets had difficulty detecting misfit when the GGUM was fit to 2PLM generated data with 15 items. In contrast, Nye et al. (2020) found that $adj\ \chi_i^2$ ratio for item doubles and triples were among the most accurate indicators of misfit; data were generated according to the 2PLM and 3PLM. However, the $adj\ \chi_i^2$ ratio for single items did not perform as well in detecting misfit as item doubles and triples. Nevertheless, power did improve for $adj\ \chi_i^2$ ratio for single items once the number of items was greater than 20. The adjusted chi-square fit statistics will be denoted as either $adj\ \chi_i^2$ ratio or ratios in the following.

The above statistic seeks to determine if the model is correct (i.e., absolute fit). An alternative approach is to determine which of a set of candidate models fits the best (i.e., relative fit). Two commonly used measures of relative model fit are the Akaike's information criterion (*AIC*, Akaike, 1973) or Bayesian information criterion (*BIC*, Schwarz, 1978). These indices penalize the log-likelihood function for the number of model parameters. *BIC* differs from *AIC* by using a penalty that also involves the sample size. *AIC* is given by $-2lnL + 2v$ and *BIC* is $-2lnL + v\ \log(N)$, where *lnL* is the log-likelihood, *v* is the number of parameters in the model, and *N* is the sample size. These relative fit statistics have shown promising

results for correctly identifying fit for ideal point and dominance models under conditions different than those studied here (Nye et al., 2020).

Contribution of the Current Study to the GGUM Model Fit Literature

In many situations in which IRT is applied the practitioner is interested in applying a particular model rather than selecting among competing models. Thus, an absolute fit statistic would be useful. The current study seeks to shed light on the mixed findings discussed above. To accomplish this objective the study's design is similar in scope to studies such as Nye et al. (2020) and Tay et al. (2011). In part, this study attempts to verify previous findings (cf. replicability crisis) while simultaneously contributing new knowledge. To this end, there are notable differences. First, no study has used the ΔAIC index with ideal point models; the ΔAIC index is discussed below. Additionally, Tay et al. (2011) used $adj\ \chi_i^2$ ratios in a *relative* fashion and software which is most likely no longer available to the practitioner. For instance, the FORSCORE software used for obtaining the $adj\ \chi_i^2$ ratios is dated 1993 and the GGUM2004 (MMLE) calibration program used was an unreleased version in 2011 (i.e., its comparability to the released version cannot be assumed). This study uses software which is freely available as R packages and because of its inclusion of *AIC* does not use $adj\ \chi_i^2$ ratios in a relative fashion, but from an absolute perspective.

Additionally, Tay et al. (2011) dropped items located in the middle of the continuum because it was "difficult to obtain IRT estimates of items close to the middle of the continuum with dominance procedures" (p. 292). In contrast, no items were dropped in this study and we did not encounter difficulties with obtaining estimates in the middle of the continuum. Finally, Tay et al. (2011) did not examine *AIC* and essentially looked at upper asymptote misfit. Our inclusion of the 3PLM allows us to examine lower tail asymptotic misfit.

In this study and with respect to *AIC*/*BIC* we introduce to the psychometric literature the use of ΔAIC and its use with screening values. In terms of the absolute fit index $adj\ \chi_i^2$ ratio tests, previous studies have assumed the conventional "critical value" of 3 for misfit detection used with dominance models is appropriate for ideal point models. This value comes from analyses involving dominance models with empirical data although it is unclear what the justification for this value is. Additionally, in the fit studies presented above the value of 3 was assumed to be applicable to the ideal point model. Moreover, no studies were found that investigated other screening values under known conditions. (Tay et al. [2011] suggests that a value of 3 be re-examined.) In the present study we examine several screening values in addition to 3 to assess the effect on fit detection. Third, previous simulation work has always used the location range [-2, 2]. This range is well within the -3 to 3 integration range used for calculating the chi-squares' expected values. As such, the [-2, 2] range does not allow an assessment of the $adj\ \chi_i^2$ ratios with values that may be observed in practice nor can past results be generalized to locations outside of the [-2, 2] range. Furthermore, extending the range captures the non-overlapping regions of the IRT models' item response functions (IRFs) which is seen as the divergence above a theta level of 2 (see Figure 1, left panel). Thus, the aforementioned proposed methodological differences between the current study and previous ones complement and contribute to the extant literature.

Methods

Factors

The four factors examined were sample size (*N*: 500, 1000, 2000, 3000), number of items (*I*: 10, 20, 40), generation model (GGUM, 2PLM, 3PLM, GRM), and screening value. Relative to the GGUM the 2PLM creates misfit in the upper asymptote, whereas the 3PLM is used to create misfit in the upper and lower asymptotes. The polytomous data are comprised of four ordered response categories. Within each condition,

data are generated and calibrated 100 times (i.e., 100 replications). For the absolute fit indices the performance for screening values between 0.25 and 3 in 0.25 increments were examined. All simulations, calibrations, and estimations of the relative model and absolute fit indices were conducted in R using `GGUM` (Tendeiro & Castro-Alvarez, 2020), `mirt` (Chalmers, 2012), and `catIrt` (Nydick, 2014) packages. MMLE execution parameters for `GGUM` and `mirt` were matched to one another. The choice of sample sizes and number of items was based on previous model fit simulation studies (Nye et al., 2020; Roberts, 2008; Tay et al., 2011) and recommended rough guidelines (see De Ayala, 2009). Nonconvergence was encountered a few times with the 10-item condition. However, in these cases the data were regenerated until a data set was obtained that led to convergence.

Data Generation

***Item and Person* Parameters.** Person parameters were randomly sampled from a $N(0, 1)$ and were allowed to vary across replications. The distributions for item parameter generation came from Nye et al. (2020), Roberts et al. (2002), and Tay et al. (2011). The item parameters were allowed to vary across replications. For the 2PLM and 3PLM the α_i were randomly sampled from a log-normal (0, 0.5) distribution and following Tay et al. (2011) divided by 1.702 with the item locations randomly sampled from a uniform distribution $U[-3, 3]$; the γ_i came from a $U[0, 0.3]$. The GGUM's α_i were randomly sampled from a $U[0.5, 2.0]$, the τ_{ik} from a $U[-1.4. -0.4]$, with δ_is randomly sampled from a $U[-3, 3]$. This range permits an examination of the effect large δs might have on the proportion of correct model identification.

For the polytomous data the α_i for the GRM were sampled from a $LN(0, 0.5)$ distribution and following Tay et al. (2011) divided by 1.702, whereas for the GGUM they came from a uniform random distribution [0.5, 2.0]. The GRM category boundary locations were randomly generated from $U[-2, -0.5]$, $U[-0.5, 0.5]$, $U[0.5, 2]$, respectively (see Nye et al., 2020). For the GGUM and following Roberts et al. (2002), the τ_{ik} were generated independently for each item. For a selected item i, the highest τ_{ik} (τ_3) was drawn from a $U[-1.4, -0.4]$. Successive τs for each item (i.e., τ_2 or τ_1) were sampled using the following recursive formula:

$$\tau_{ik-1} = \tau_{ik} - 0.25 + e_{ik-1}, \text{ for } k = 2, 3, \ldots, F, \quad (7)$$

where e_{ik-1} represents a random error term sampled from a $N(0, 0.04)$, F is the number of observable response categories minus 1. The item δ_is were randomly sampled from $U[-3, 3]$.

Response Data Generation. Dichotomous responses were generated by comparing a model's item response probabilities to a $U[0, 1]$. A response of 1 was assigned if the uniform random was less than the item probability, 0 otherwise.

For polytomous data ($m_i = 3$) the sums of successive response probabilities for categories 0, 1, and 2 were obtained and compared to a random number from $U[0,1]$. A response (score) of 0, 1, or 2, was assigned if the randomly sampled uniform number was less than or equal to the category 0 probability, the sum of category probabilities 0 and 1, or the sum of category probabilities 0, 1, 2, respectively, otherwise the response was a 3. For the dominance models `catIrt`'s 1-based responses were recoded to be 0-based. For the GGUM model response data are generated using the `GGUM` R package by setting the program's category threshold indicator to 1 and 3 for dichotomous data and polytomous data, respectively (Tendeiro & Castro-Alvarez, 2020).

Model Calibration

The GGUM model was fit to the data using the `GGUM` package (Tendeiro & Castro-Alvarez, 2020); `GGUM` uses MMLE. The selected number of nodes, maximum iterations, and convergence tolerance values followed those in Tay et al. (2011). The 2PLM, 3PLM, and GRM models were fit to the data using `mirt` and to obtain the relative fit indices. With `mirt` MMLE was selected for model calibration and execution

parameters (e.g., the number of nodes, maximum iterations) were matched to those of GGUM. Descriptive statistics for the 3PLM showed that the median γ_i ranged from 0.06 to 0.14. However, because the corresponding the mean γ_i ranged from 0.14 to 0.19 `mirt` had difficulty estimating γ_i for one or more items. The difficulty was most pronounced for the $N = 500/I = 10$ condition with a difference between the *M* and median of 0.13 and, as one would expect, became progressively smaller as *N* and *I* increased. Thus, the use of priors was implanted in phases.

In phase 1 beta and normal with different shape or location/variability parameters, respectively, were utilized to improve estimation of γ_i. Beta prior results were comparatively poor with `mirt` providing the warning "Lower and upper bound parameters (g and u) should use 'norm' (i.e., logit) prior". Of the three normal priors examined a normal prior with $M = -1.5$ and $SD = 0.5$ produced the best results.

Descriptive statistics by condition showed that `mirt` had difficulty estimating the discrimination parameter(s) for one or more items. Consequently, in phase 2 a series of lognormal priors for discrimination estimation was investigated. The results showed a $LN(0.2, 0.2)$ prior worked best.

Because descriptive statistics for each condition showed one or more items with location estimates that were in the double digits, phase 3 examined different priors for the intercept. Corresponding results led to a normal prior located at 0 with variability 1.5 being selected.

Adjusted Chi-square (Absolute Model Fit index)

Because for the absolute fit index model comparisons are unnecessary only the GGUM was fitted to each data generation model. As one would do in practice the $adj\ \chi_i^2$ ratios were calculated using the estimated parameters in each replication were used. To compare the performance of the item-level fit statistics ($\chi 2$: item singles) to item subsets-level fit statistics (e.g., $adj\ \chi_i^2$ ratios based on item subsets for item doubles and triples) the proportion of items exhibiting misfit per replication were averaged across replications.

Evaluating model fit for the average $adj\ \chi_i^2$ ratios for item singles, doubles, and triples entailed dividing each index by their corresponding *df*. Ratios greater than a given screening value of, for example, 3 indicate misfit. The corresponding correct detection rates across replications for each of the fit indices were calculated as an indicator of model fit/misfit. For example, if the fit statistic led to incorrectly rejecting the model-data fit hypothesis when the GGUM model is fit to GGUM generated data four times across the 100 replications, then the incorrect detection rate is .04. Conversely, if the GGUM model is fit to the 2PLM generated data and the fit statistic led to correctly rejecting the model-data fit hypothesis 89 times across the 100 replications, then the correct detection rate is .89. Unlike other studies (e.g., Nye et al., 2020) that use, for example, the term "power" and Type I error rate we use the term "correct detection" and "incorrect detection" (proportion) rates, respectively. Our reasoning lies in the number of replications (100). Specifically, 100 replications are statistically insufficient for us to treat our proportions as reflective of probabilities. Rather our proportions are indicative of the relative frequency one might observe over the long run (i.e., a probability). Unfortunately, the execution times with current computing power and software does not realistically allow performing 10,000 replications or more to obtain an accurate estimate of a probability.

Relative Model Fit Indices

In contrast to above, to assess relative fit each model was fitted to each data generation model. Specifically, for the dichotomous condition the GGUM, 2PLM, and 3PLM were fit to each data generating models. Similarly, for the polytomous condition the GGUM and GRM were fit to each data generating models. *AIC* was utilized using two conventions. First, the model with the lowest AIC_{min} (or BIC_{min}) was selected as the "best" fitting model. Second, the difference between model *d*'s *AIC* and the minimum *AIC*

was calculated ($\Delta AIC = AIC_d - AIC_{min}$). Following Burnham, Anderson, and Huyvaert (2011) a $\Delta AIC \leq 2$ indicates that model *d* shows substantial support relative to the model with the minimum *AIC* and a $4 \leq \Delta AIC \leq 7$ shows that model *d* has some support; a $\Delta AIC \geq 10$ shows no support for model *d*. For example, let AIC_{min} be for the GGUM and AIC_k for the 2PLM, then a $\Delta AIC \leq 2$ indicates strong evidence in favor of the 2PLM and a $4 \leq \Delta AIC \leq 7$ indicates some support for the 2PLM fitting as well as the GGUM. Across the 100 replications the number of times a model was selected as the best fitting was recorded for each condition.

Results

Absolute Fit Indices for Dichotomous Data

Figure 2 presents the correct detection proportions for $adj\ \chi_i^2$ ratios for item singles, doubles, and triples across the 100 replications as a function of screening value. As can be seen, correct detection rate is affected by *I*, *N*, and item variants (i.e., singles, doubles, triples) to a greater extent at the lower end of screening value scale than at its upper end. For screening values of 1.0 and larger and for all item variants, the average $adj\ \chi_i^2$ ratio statistics show correct detection rates greater than .90 and typically closer to 1.0 when the GGUM was the underlying data model. Correct detection rates for ideal point data progressively fell as the screening value decreased, this rate of decline varied with item variant, *N*, and *I* except for singles with a $N = 500$. These declines were associated with an increase in correct detection of dominance data (i.e., 2PLM, 3PLM) with, generally speaking, larger *N* associated with greater improvement in detection by item doubles and triples. For $I = 10$ increasing *N* led to an improvement in detecting that GGUM was inappropriate for the 2PLM (and to a lesser extent the 3PLM). When $N < 1000$ increasing the instrument length leads to a comparative improvement in correct detection of dominance data albeit proportions are less than .5. When $I = 20$ or 40 and $N = 3000$ there is an increase in the correct detection of the 2PLM data (and to a lesser extent the 3PLM data) with a concomitant decrease in correctly identifying ideal point data.

Contrasting Figures 2a, 2b, and 2c shows that for dominance data item doubles and triples outperformed item singles in correct detection rates for screening values less than approximately 1.0. For $N = 3000$ and screening values 0.5 or less the item doubles correct detection of dominance data ranged from .3 to .87, but correct detection of ideal point data decreased from 0.3 to .65. For $N < 2000$ item triples correct detection of dominance data improved with increasing I with correct detection of the 2PLM approaching .8. When $N = 3000$ and a screening value 0.25 correct detection of dominance data was effectively 1.0, but 0 with respect to ideal point data. When the screening value increased to 0.5, then correct detection of dominance data fell to 60% to 65% as was the case with ideal point data. This screening value is far below the typically used value of 3.0.

Relative Fit Indices for Dichotomous Data

Table 1 presents the proportion of times a model was selected as best fitting by AIC_{min} or by ΔAIC, and the median ΔAIC. (Because of space limitations we present only the *AIC* results; *BIC* results are similar and are available from the first author). The GGUM was correctly identified by AIC_{min} and ΔAIC as best fitting the GGUM data 100% of the time. For the 2PLM data AIC_{min} and ΔAIC correctly identified the appropriate model over 95% of the time when $I \geq 20$ and $N \geq 500$. However, with $I = 10$ AIC_{min} and ΔAIC incorrectly identified GGUM as the appropriate model with proportions from .64 to .99 contingent on *N*. Neither AIC_{min} and ΔAIC performed well with the 3PLM generated data.

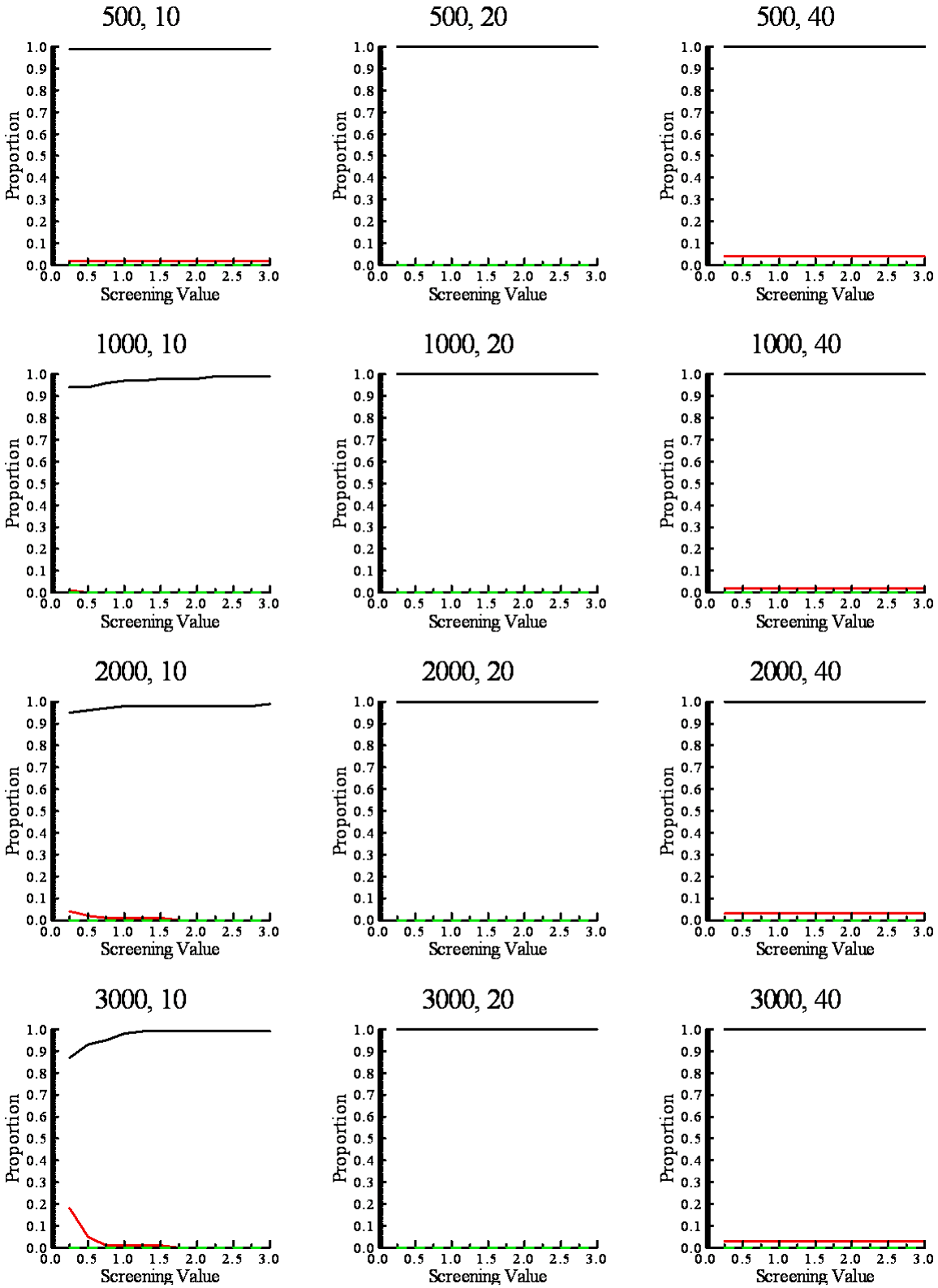
Absolute Fit Indices for Polytomous Data

Figure 3 presents the correct detection proportions for $adj\ \chi_i^2$ ratios for item singles, doubles, and triples across the 100 replications as a function of screening value. As can be seen, for screening values 1 and above and for all item variants, the average $adj\ \chi_i^2$ ratio statistics showed correct detection rates between .96 and 1.00 when the GGUM was the underlying data model. In contrast, the average $adj\ \chi_i^2$ ratio statistics (singles) correctly detected that the GGUM was a "misfit" (i.e., the underlying model was the GRM) increased as the screening value decreased without a concomitant decrease correct detection rate of ideal point data. This pattern did not hold for item doubles and triples. In contrast to what is seen with dichotomous data, with item doubles and triples there was a "sweet spot" that maximize correct detection of both ideal point data and dominance data; both greater than .95 correct detection rate. This screening value shifted up the scale from approximately 0.5 for $N = 500$ to about 1.0 when $N = 3000$. In general, for these polytomous data the screening value 0.75 might be considered to be a good compromise across the N and I conditions. This screening value could also be considered to work reasonably well with item singles because the correct detection rate would exceed 80% for both ideal point and dominance data except for when $N = 500$ and $I = 10$. As was the case with dichotomous data, this screening value is far below the typically used value of 3.0.

Relative Fit Indices for Polytomous Data

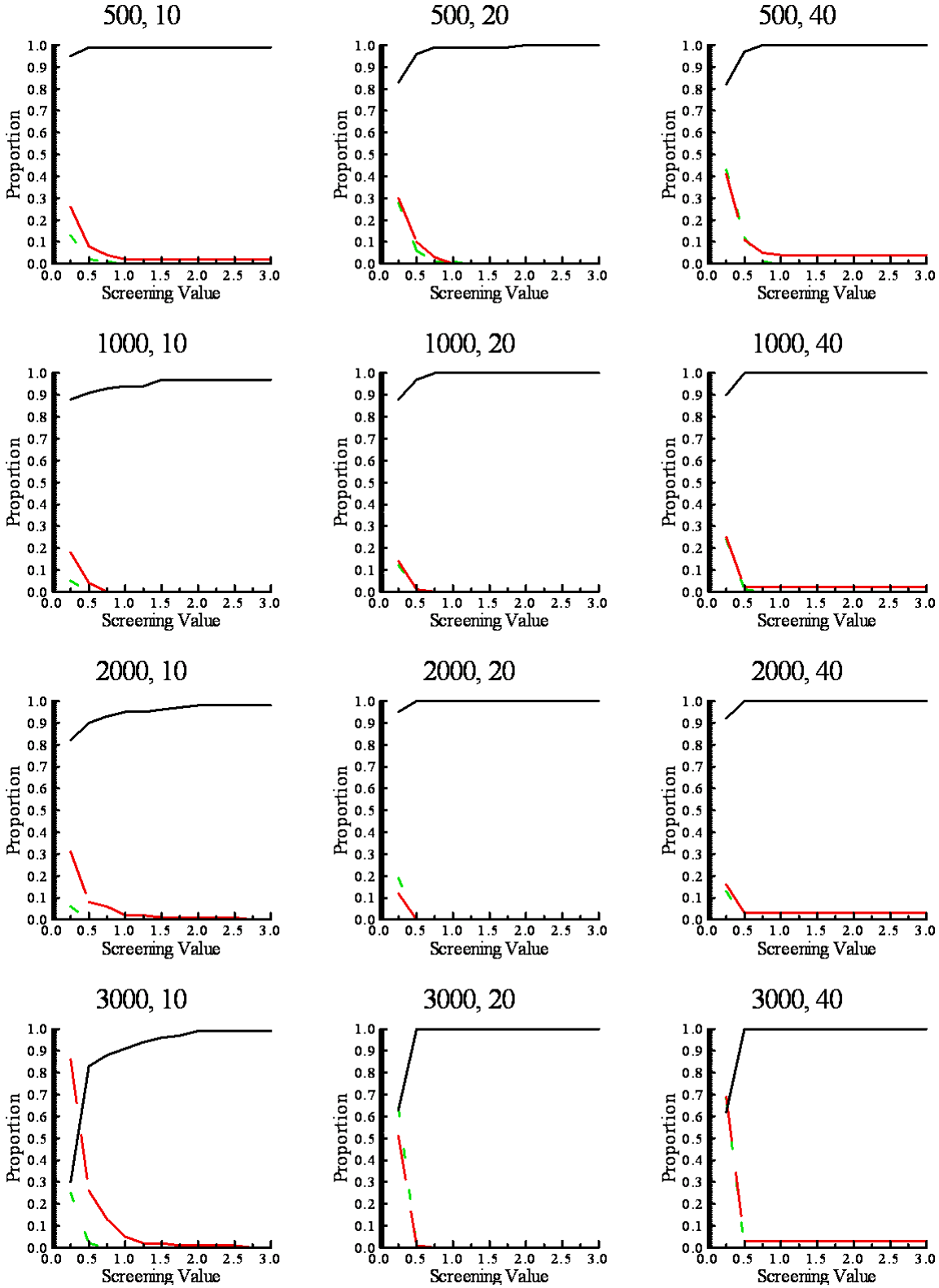
Table 2 presents the proportion of times a model was selected as best fitting by AIC_{min} or ΔAIC, and the median ΔAIC for the GGUM and GRM polytomous data. Using AIC_{min} or ΔAIC the GGUM model was found to best fit the GGUM data regardless of the number of items and sample size 100% of the time. Moreover, for the GRM data and when $I \geq 20$ AIC_{min} and ΔAIC correctly identified the GRM greater than 99% of the time regardless of N. With less than 20 items AIC_{min}'s and ΔAIC's accuracy decreased, but still tended to correctly identify the GRM the majority of the time with AIC_{min} and ΔAIC showing comparable results.

Figure 2a. $adj\ \chi_i^2$ ratios (singles): Correct detection proportions (dichotomous) vs. screening values, *I*, *N*.



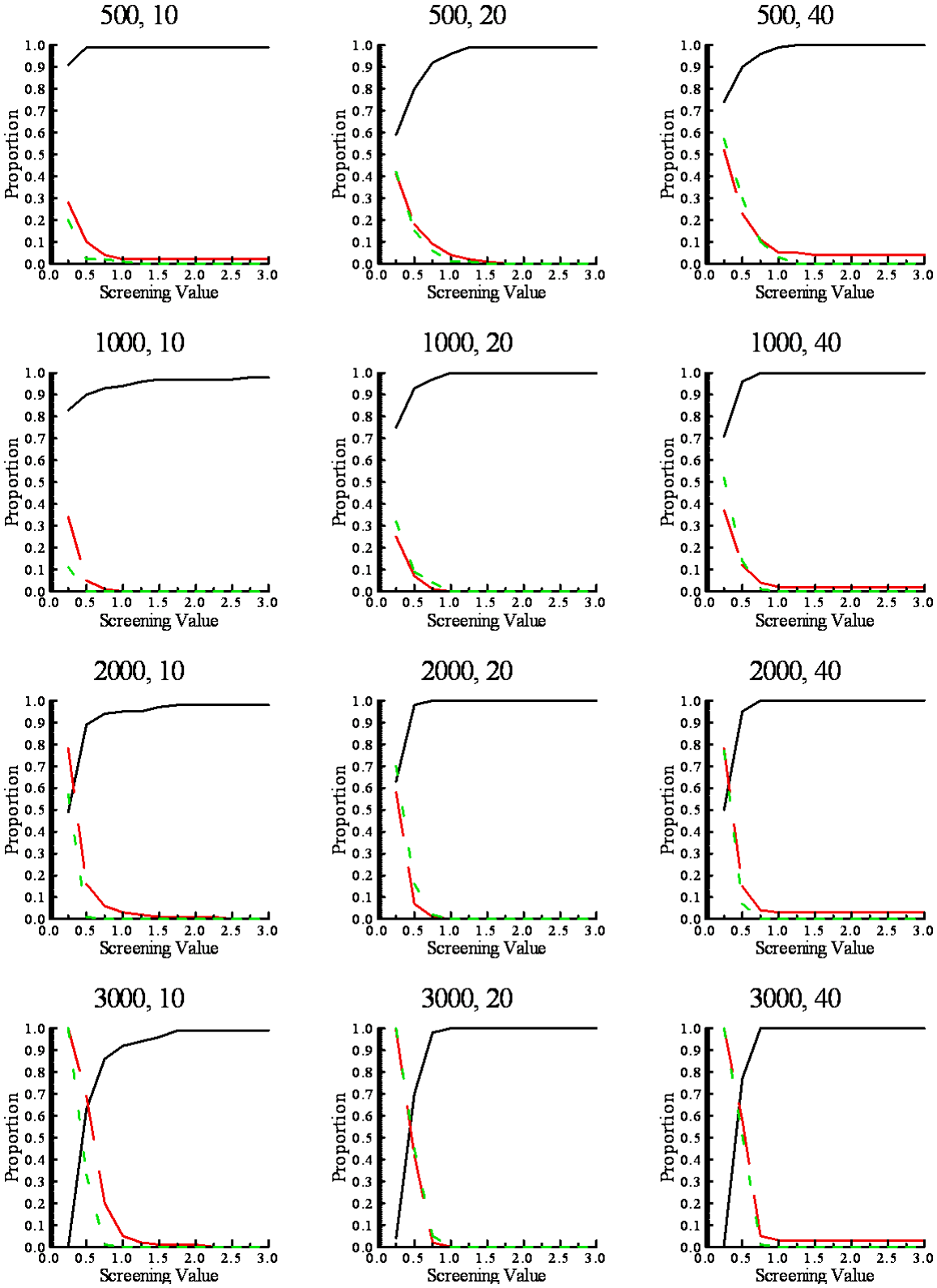
Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red large dash line is 2PLM, green dash line is 3PLM.

Figure 2b. $adj\ \chi_i^2$ ratios (doubles): Correct detection proportions (dichotomous) vs. screening values, *I*, *N*.



Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red large dash line is 2PLM, green dash line is 3PLM.

Figure 2c. $adj\ \chi_i^2$ ratios (triples): Correct detection proportions (dichotomous) vs. screening values, *I*, *N*.



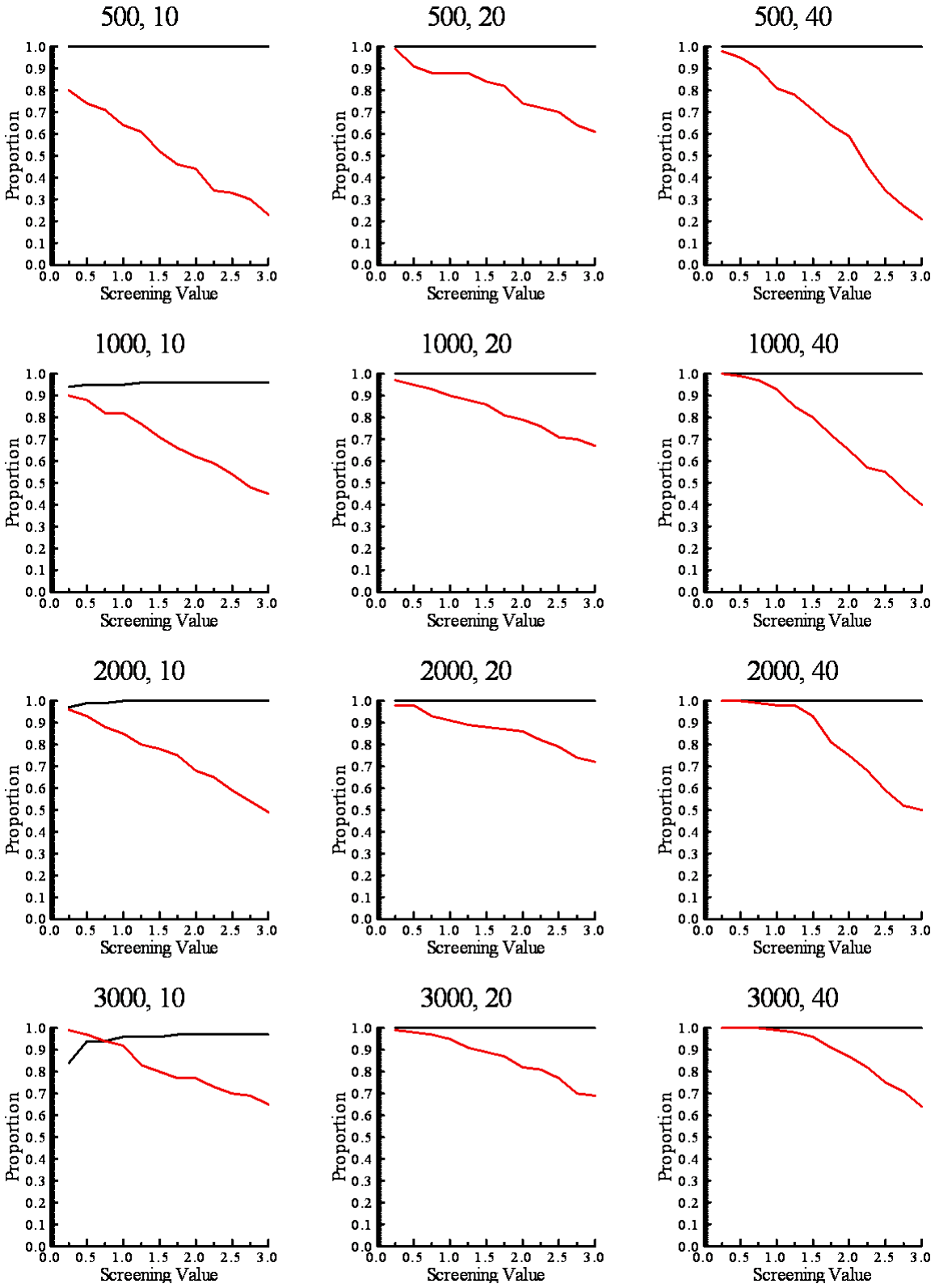
Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red large dash line is 2PLM, green dash line is 3PLM.

Table 1. Relative Fit Indices correct and incorrect detection proportions, dichotomous data.

		Model	GGUM Data			2PLM Data			3PLM Data		
I	*N*	Selected	*AIC*	*ΔAIC*	*ΔAIC<7*	*AIC*	*ΔAIC*	*ΔAIC<7*	*AIC*	*ΔAIC*	*ΔAIC<7*
10	500	GGUM	1	81.43	1	0.64	31.72	0.66	0.2	5.98	0.27
		2PLM	0	-	0	0.36	10.49	0.34	0.8	11.95	0.73
		3PLM	0	-	0	0	-	0	0	-	0
	1000	GGUM	1	163.41	1	0.85	50.44	0.85	0.43	9.78	0.50
		2PLM	0	-	0	0.15	10.45	0.15	0.57	9.87	0.50
		3PLM	0	-	0	0	-	0	0	-	0
	2000	GGUM	1	302.73	1	0.96	116.25	0.96	0.58	25.27	0.59
		2PLM	0	-	0	0.04	13.28	0.04	0.42	13.46	0.41
		3PLM	0	-	0	0	-	0	0	-	0
	3000	GGUM	1	501.59	1	0.99	171.07	0.98	0.68	37.10	0.72
		2PLM	0	-	0	0.01	6.03	0.02	0.32	15.48	0.28
		3PLM	0	-	0	0	-	0	0	-	0
20	500	GGUM	1	187.41	1	0.01	17.77	0.01	0	-	0
		2PLM	0	-	0	0.99	31.57	0.99	1	34.92	1.00
		3PLM	0	-	0	0	-	0	0	-	0
	1000	GGUM	1	330.67	1	0.04	21.16	0.04	0	-	0
		2PLM	0	-	0	0.96	33.11	0.96	1	35.86	1.00
		3PLM	0	-	0	0	-	0	0	-	0
	2000	GGUM	1	642.48	1	0.05	33.20	0.04	0	-	0
		2PLM	0	-	0	0.95	31.84	0.96	1	35.56	1.00
		3PLM	0	-	0	0	-	0	0	-	0
	3000	GGUM	1	1058.77	1	0.05	42.91	0.13	0	-	0
		2PLM	0	-	0	0.95	32.39	0.87	1	35.44	1.00
		3PLM	0	-	0	0	-	0	0	-	0
40	500	GGUM	1	460.76	1	0	-	0	0	-	0
		2PLM	0	-	0	1	64.17	1	1	66.24	1
		3PLM	0	-	0	0	-	0	0	-	0
	1000	GGUM	1	957.88	1	0	-	0	0	-	0
		2PLM	0	-	0	1	65.81	1	1	66.93	1
		3PLM	0	-	0	0	-	0	0	-	0
	2000	GGUM	1	1958.87	1	0	-	0	0	-	0
		2PLM	0	-	0	1	65.37	1	1	65.61	1.00
		3PLM	0	-	0	0	-	0	0	-	0
	3000	GGUM	1	2977.17	1	0	-	0	0	-	0
		2PLM	0	-	0	1	66.38	1	1	63.59	1.00
		3PLM	0	-	0	0	-	0	0	-	0

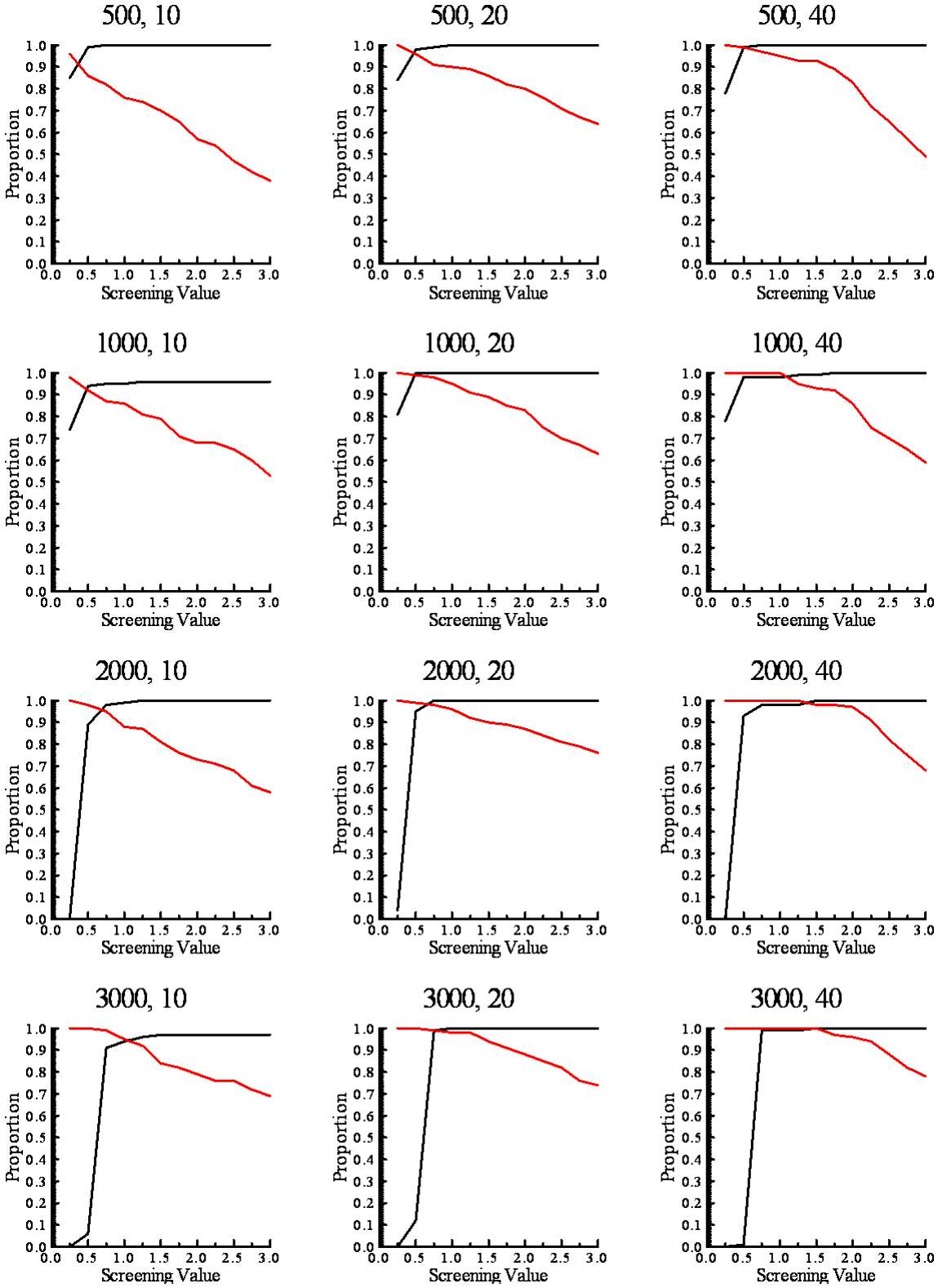
Data generation model indicated by "<model> + Data"; ⊥AIC: median value across replications; AIC is AIC_{min}. For AIC and ⊥AIC<7 shaded cells indicate correct proportion matches between fitted model and data generating model and unshaded cells indicate incorrect proportion matches between fitted model and data generating model (i.e., lowest AIC and/or ⊥AIC<7 obtained by a mismatching model relative to the model's generated data). Priors used with 3PLM.

Figure 3a. $adj\ \chi_i^2$ ratios (singles): Correct detection proportions (polytomous) vs. screening values, *I*, *N*.



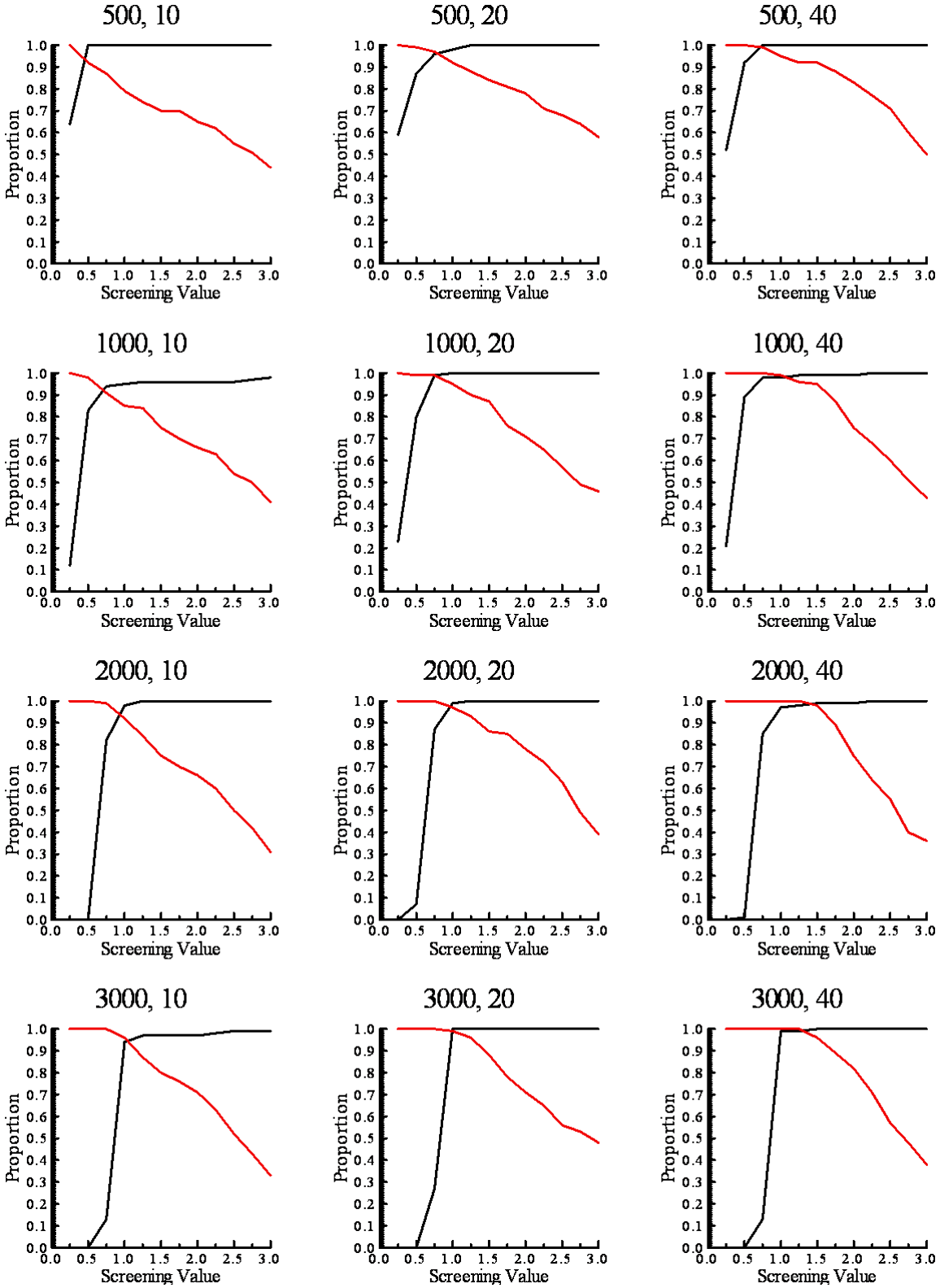
Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red line is GRM.

Figure 3b. $adj\ \chi_i^2$ ratios (doubles): Correct detection proportions (polytomous) vs. screening values, *I*, *N*.



Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red line is GRM.

Figure 3c. $adj\ \chi_i^2$ ratios (triples): Correct detection proportions (polytomous) vs. screening values, *I*, *N*.



Plots labelled in terms of N and I (e.g., '500, 10': N = 500, I = 10). Data generation model: Black line is GGUM, red line is GRM.

Table 2. Relative Fit Indices correct and incorrect detection proportions, polytomous data.

			GGUM Data			**GRM Data**		
I	*N*	Model Selected	*AIC*	*ΔAIC*	*ΔAIC<7*	*AIC*	*ΔAIC*	*ΔAIC<7*
10	500	GGUM	1	374.37	1	0.28	32.92	0.27
		GRM	0	-	0	0.72	88.05	0.73
	1000	GGUM	1	745.33	1	0.33	56.49	0.27
		GRM	0	-	0	0.67	192.97	0.73
	2000	GGUM	1	1378.61	1	0.34	169.90	0.29
		GRM	0	-	0	0.66	293.91	0.71
	3000	GGUM	1	2278.15	1	0.33	257.39	0.32
		GRM	0	-	0	0.67	389.72	0.68
20	500	GGUM	1	838.04	1	0.01	2.87	0
		GRM	0	-	0	0.99	290.52	1
	1000	GGUM	1	1791.31	1	0	-	0
		GRM	0	-	0	1	536.59	1
	2000	GGUM	1	3579.41	1	0	-	0
		GRM	0	-	0	1	976.07	1
	3000	GGUM	1	5195.68	1	0	-	0
		GRM	0	-	0	1	1485.43	1
40	500	GGUM	1	2074.6	1	0	-	0
		GRM	0	-	0	1	420.76	1
	1000	GGUM	1	4280.06	1	0	-	0
		GRM	0	-	0	1	751.9	1
	2000	GGUM	1	8307.04	1	0	-	0
		GRM	0	-	0	1	1400.44	1
	3000	GGUM	1	12,763.51	1	0	-	0
		GRM	0	-	0	1	2203.92	1

Data generation model indicated by "<model> + Data"; 4 response categories; ⊿AIC: median value across replications; AIC is AIC_{min}. For AIC and ⊿AIC<7 shaded cells indicate correct proportion matches between fitted model and data generating model and unshaded cells indicate incorrect proportion matches between fitted model and data generating model (i.e., lowest AIC and/or ⊿AIC<7 obtained by a mismatching model relative to the model's generated data).

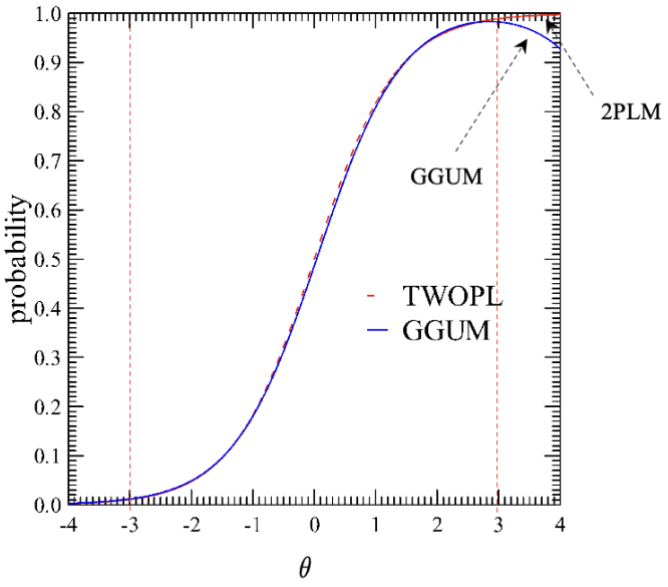
Discussion

This study examined the performance of the absolute $adj\ \chi_i^2$ ratio statistics and relative fit with AIC_{min} and ΔAIC indices for the GGUM with dichotomous and ordered polytomous dominance and ideal point data. As indicated in the Contribution of the Current Study section, the objective was to contribute to the model-data fit work in this area by verifying previous findings, more fully investigating the screening value issue, and resolving previous conflicting results.

For the absolute fit statistic results and item singles, doubles, and triples the GGUM fit the GGUM data correct detection rate increased as screening value increased with a detection rate of over 96% of the time with screening value of 0.75 or higher. Therefore, if the practitioner/researcher believes that an individual will agree with, for example, an attitude statement that reflects the participant's view (i.e., $\delta_i \approx \theta$) and will disagree with a statement that is less or more extreme than their view (i.e., ideal point data), then this result is encouraging. In contrast, with dominance data the $adj\ \chi_i^2$ ratio test correctly detected that the GGUM was a "misfit" only between 0% and 20% performing slightly better with the 2PLM than with the 3PLM data; this result is similar to Tay et al. (2011). Recall that with dominance data the practitioner/researcher believes that an individual will agree with a statement to the extent that the statement reflects a perspective that is equal to less extreme than their view (i.e., for $\delta_i < \theta$ then $p(x = 1)$ is maximized).

In certain situations $adj\ \chi_i^2$ ratio's performance can be explained by noting that its expected value assumes a unit normal distribution with a quadrature integration range of [-3, 3]. For example, assume 2PLM data are generated for an item located at 0.0 (Figure 4). When the GGUM is fit to these data its item location is estimated to be higher, say 2.85, while simultaneously adjusting α_i and τ_{ik} to shift the modal probability up the scale and to broaden and increase its height so there is little discrepancy between the two IRFs between -3 and 3; examination of various misfitting items verified that this occurred. The $adj\ \chi_i^2$ ratio reflects this correspondence between -3 and 3 and the GGUM is found to fit the 2PLM data below 3. Although the GGUM's upper asymptote for this item is 0 the IRF's transition to a monotonically decreasing function occurs above 3. Thus, the 2PLM and GGUM IRFs' upper asymptotes discrepancy occurs above the integration range's upper bound and is not reflected in the expected value. Thus, the IRFs discrepancy that would distinguish the two models from one another is not captured by the $adj\ \chi_i^2$ ratio. Similarly, with the 3PLM lower asymptote the IRFs the discrepancy (i.e., between an IRF with a $\gamma_i = 0.2$ and the GGUM lower asymptote of 0) may occur below -3 and would not be reflected in the $adj\ \chi_i^2$ ratio (This explanation generalizes to item doubles and triples.) Because the GGUM can fit the 2PLM (or 3PLM) data between -3 and 3, the context determines if fit in this interval is good enough for the intended purpose. By increasing the integration interval to [-4, 4] or [-6, 6] and a non-unit normal the $adj\ \chi_i^2$ ratio's ability to distinguish between upper and lower asymptotes could improve.

Figure 4. Response Functions (GGUM, $\alpha = 1.45$, $\delta = 2.85$, $\tau 1 = -2.8$; 2PLM, $\alpha = 1.5$, $\delta = 0.0$)



Red dash lines depict lower/upper bounds of integration range for E_i (k,k').

With ordered polytomous data we see different $adj\ \chi_i^2$ ratio results. With these data the GGUM will have some ORFs that are asymptotic with 0.0. Because, conditional on theta, the sum of the probabilities of responding to the item options is 1.0, then there is one ORF that must be asymptotic with 1.0 (see Figure 1 right panel). By definition this ORF will not transition to a monotonically decreasing function as θ approaches $+\infty$ as the IRF does in the dichotomous case. Thus, the $adj\ \chi_i^2$ ratio can distinguish between the GRM and GGUM by the correspondence between the models' ORFs within the [-3, 3] range.

For all intents and purposes reducing the screening value for the $adj\ \chi_i^2$ ratio did not have a meaningful impact in detecting GGUM misfit with the dichotomous dominance data unless one considers correct detection rates of 60% to 65% acceptable. However, with ordered polytomous data we see that instrument length, sample size, and screening value interacting to enhance the $adj\ \chi_i^2$ ratio correct detection rate for identifying the GGUM not fitting the GRM data while still maintaining the ability to correctly identify ideal point data regardless of whether item singles, doubles, and triples are used. Moreover, it appears that part of the performance differences between item singles, doubles, and triples previously seen in the literature is due to using a screening value of 3. Nevertheless, results exhibit the previously seen pattern in which doubles and triples outperform item singles just not as dramatically.

In terms of the relative fit indices, dichotomous and ordered polytomous data, AIC_{min} and ΔAIC always correctly identified the GGUM as the best fitting model to the GGUM data regardless of number of items and *N*. With respect to the 2PLM and GRM data, AIC_{min} and ΔAIC correctly identified that the GGUM was not the best model at least 95% or more of the time when $I > 20$. For the dichotomous data neither *AIC* detection approach was able to correctly identify the dominance generating model when $I = 10$. Thus, with dichotomous data one sees that *AIC*/*BIC*'s utility affected by the reduction in the number of model parameters compared to the longer lengths. Stated another way, the penalties imposed by *AIC*/*BIC* do not always sufficiently compensate for model complexity. (With the polytomous data this is less of an issue because of the increase in the number of model parameters.) Contrasting AIC_{min} with ΔAIC one sees very little difference in performance.

Recall that *AIC*, ΔAIC, and *BIC* do not specify whether a model fits the data, but only that in comparison to candidate models a particular model fits the best. For example, the 2PLM may be selected as fitting a two-dimensional data set with an interdimensional correlation of .05 better than the 1PLM or 3PLM not that the 2PLM is the true model nor that it fits the data. As seen in this study, the 2PLM fit the 3PLM data better than the 3PLM.

Utilizing absolute fit diagnostic item-level information would allow one to determine for which item, if any, the model is not functioning well and thereby permit the practitioner to make appropriate modifications. Using $adj\ \chi_i^2$ ratio with a modified integration range and a focus on item and not model-level fit (as done above) might be useful for this purpose.

This study's $adj\ \chi_i^2$ ratio test results do not fully agree with those from previous studies with respect to item doubles and triples. This discrepancy may be due to methodological differences between this study and that of Tay et al. (2011). In Tay et al. (2011), the assessment of fit examined the $adj\ \chi_i^2$ ratios between IRT models in relative terms. That is, the proportion of the $adj\ \chi_i^2$ ratio values across replications obtained by fitting the correct model to its data was compared to those of a misspecified model (i.e., the comparison was relative to the misspecified model). In contrast, in this study the performance of the $adj\ \chi_i^2$ ratios for item singles, doubles, and triples involved comparing the $adj\ \chi_i^2$ ratios screening values to indicate fit; this is akin to practice. Thus, this difference might account for the more favorable results found in Tay et al. (2011).

General Guidelines for the GGUM Fit Assessment

If one intends to utilize the $adj\ \chi_i^2$ ratios or relative fit indices to assess fit for the GGUM with either empirical data (e.g., from noncognitive measures) or simulated data we present general guidelines for consideration. First, if one fits the GGUM to empirical data, then the $adj\ \chi_i^2$ ratios tests of absolute fit will almost always perform better at correctly detecting misfit with polytomous data than with dichotomous data. Thus, personality statements with a dichotomous response format that are, theoretically, best represented by ideal point models should not rely solely on the $adj\ \chi_i^2$ ratios tests of absolute fit. Second, the cutoff mean ratio value of 3 should not be taken for granted as valid for model-level fit assessment applications. Rather, absolute fit diagnostic item-level information might be a better option, specifically for dichotomous data. Moreover, although at the model-level a cutoff mean ratio value of 0.75 for polytomous data seems promising, additional research should examine its applicability under additional conditions (e.g., a different number of options, with unordered polytomous data). Finally, because with dichotomous data the penalties imposed by *AIC/BIC* do not always sufficiently compensate for model complexity at short instrument lengths their use by researchers should be done with care.

Received: 9/19/2023. **Accepted:** 11/17/2025. **Published:** 12/04/2025.

Citation: Alzarouni, A, & De Ayala, R. J. (2025). Assessing model fit of the generalized graded unfolding model. *Practical Assessment, Research, & Evaluation*, 30(1)(10). Available online: https://doi.org/10.7275/pare.2044

Corresponding Author: Abdulla Alzarouni, University of Nebraska-Lincoln[1]
Email: abdulla_alzarouni@yahoo.com

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Ames, A., & Penfield, R. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, *34*, 39-48.

Arpaly, N., & Schroeder, T. (2018). *In praise of desire* (Reprint, Vol. 1, pp. 18–42). eBook: Oxford University Press.

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

Burnham, K.P., Anderson, D.R., & Huyvaert, K.P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65, 23-35.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116–131.

[1] Now at The University of Texas at Austin. UT Austin email: aa223957@my.utexas.edu

Chalmers, R. P. (2012). `mirt`: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Chernyshenko, O., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523-562.

Chernyshenko, O., Stark, S., Drasgow, F., & Roberts, B.W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88-106.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

Drasgow, F., Chernyshenko, O., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology*, *3*, 465-476.

Drasgow, F., Levine, M., Tsien, S., Williams, B., & Mead, A. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143-166.

Nydick, S. W. (2014). `catIrt`: An R package for simulating RIT-based computerized adaptive Tests. R package version 0.5-0. https://CRAN.R-project.org/package=catIrt

Nye, C., Joo, S., Zhang, B., & Stark, S. (2020). Advancing and evaluating IRT model data fit indices in organizational research. *Organizational Research Methods*, *23*, 457-486.

Roberts, J. (2008). Modified likelihood-based item fit statistics for the generalized graded Unfolding Model. *Applied Psychological Measurement*, *32*, 407-423.

Roberts, J., & Laughlin, J. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20*, 231-255.

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

Roberts, J., Donoghue, J., & Laughlin, J. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement*, *26*(2), 192-207.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point models be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25-39.

Tay, L., Ali, U., Drasgow, F., & Williams, B. (2011). Fitting IRT models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models. *Applied Psychological Measurement*, *35*, 280-295.

Tendeiro, J. N., & Castro-Alvarez, S. (2020). `GGUM`: Generalized graded unfolding model. R package version 0.4-1. https://CRAN.R-project.org/package=GGUM