A COMPREHENSIVE EVALUATION OF FIT STATISTICS TO IDENTIFY THE CORRECT ITEM RESPONSE PROCESS

by

LI GUAN

(Under the Direction of Nathan T. Carter and Gary J. Lautenschlager)

ABSTRACT

The psychological response process is the judgement process that individuals undergo when responding to stimuli such as self-report items. The dominance assumption – the most common assumption about response processes – implies that individuals with higher standing on the trait continuum are more likely to endorse the item. On the other hand, the ideal point assumption implies that an individual is more likely to agree with the item when it is close to their standing on the trait continuum. Although the dominance assumption has been pervasive in psychological research, new evidence suggests the ideal point model may be more appropriate for self-report measures of individual differences, such as personality, attitudes, and interests. One major challenge faced by researchers is correctly understanding the response process to enable accurate theoretical and statistical inferences about psychological attributes. Inappropriately applying a dominance scoring to ideal point data can result in major substantive misinterpretations, including finding spurious dimensions in truly unidimensional data (Davison, 1977), and can lower power in tests for curvilinearity and interactions (Carter et al., 2017).

Alternatively, inappropriate application of ideal point scoring to dominance data can result in high false positive rates for tests of curvilinearity and interactions (Carter et al., 2017). Currently, however, there is little guidance available on how best to determine whether responses arise from a dominance or ideal point response process, and the guidance that does exist is mixed. In two Monto Carlo simulation studies, the effectiveness of log-likelihood (LL), Akaike information criteria (AIC), Bayesian information criteria (BIC), M_2 statistics, root mean square error of approximation (RMSEA), Tucker-Lewis index (TLI), comparative fit index (CFI), and adjusted χ^2/df ratios in identifying the correct item response process. Even though results showed that no one universal index can point to the correct response process 100% of the time, it is more appropriate to use the LL, AIC, and BIC to assist with such decisions.

INDEX WORDS: IRT, item response process, dominance, ideal point, model-data fit statistics

A COMPREHENSIVE EVALUATION OF FIT STATISTICS TO IDENTIFY THE CORRECT ITEM RESPONSE PROCESS

by

Li Guan

M.Sc., University of Georgia, 2015

A Dissertation Submitted to the Graduate Faculty of the University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2018

A COMPREHENSIVE EVALUATION OF FIT STATISTICS TO IDENTIFY THE CORRECT ITEM RESPONSE PROCESS

by

LI GUAN

Major Professor:	Nathan T. Carter
	Gary J. Lautenschlager
Committee:	Dorothy R. Carter
	Zhenqiu Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
December 2018

ACKNOWLEDGEMENTS

First, I would like to thank my major advisor, Dr. Gary Lautenschlager, for recruited me six years ago. Also, Dr. Nathan Carter for his great support and helpful guidance. You have been always a nice, professional and knowledgeable advisor. I have learnt so much from you. Also, I would like to thank all my committee members for their critical suggestions and mentorship. Finally, I would like to thank my family for their support and their never-ending love.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

1 INTRODUCTION

Item response theory (IRT) (Lord, 1999), also known as modern test theory, has been widely used to measure cognitive ability and academic achievement over the years. In IRT models, an individual's latent trait score is estimated not only based on item responses, but also on items' idiosyncratic properties (e.g., item difficulty, item discrimination), resulting in more accurate trait score estimations than the traditional sum-score approach would produce. Furthermore, IRT can significantly benefit psychological measurement. For example, shorter tests can achieve higher reliability than longer tests; the reliability of measurement differs depending on the interaction between persons' ability levels and the administered test; and meaningful comparisons of individuals can be made when test scores are obtained across multiple tests (Embretson & Reise, 2000). Given these advantages of incorporating IRT models into psychological measurement, numerous such models have been developed to reflect diverse testing formats.

Most IRT models assume that respondents respond to items in a way that is consistent with a *dominance* response process, which implies that individuals with higher standing on the trait continuum are more likely to endorse the item. In ability tests, this indicates that people try to "dominate" or "beat" the item. For example, the popular two-parameter logistic (2PL) model (Birnbaum, 1968) used for scoring dichotomous items, and its polytomous counterpart, Samejima's graded response model (GRM; 1969), were developed to reflect the dominance

assumption; both models are grounded in ability testing[1]. Notably, each of these models assumes *monotonicity* in the relation between latent traits and the item scores, such that as the latent trait goes up, so does the probability of selecting the "correct" option or highest score on the item.

Given the benefits of using IRT models, dominance-based models have been (perhaps mistakenly) adopted by psychologists for assessing individuals' non-cognitive psychological attributes, such as personality (Tellegen, 1982), opinions (Guttman, 1950), and leadership behaviors (Zagorsek, Stough, & Jaklic, 2006). However, some researchers indicate that the more appropriate assumption in such tests is an *ideal point* response process, which theorizes that item endorsement only occurs when individuals' standing on the trait continuum is close to the level of the trait expressed by item content (Davison, 1977; Thurstone, 1927)(Thurstone, 1927; Coombs, 1964; Davison, 1977). In ideal point models, item endorsement declines as individuals' standing on the trait continuum lies further away from the item location. Researchers have demonstrated the superior fit of ideal point IRT models to non-cognitive data such as personality (Stark, Chernyshenko, Drasgow, & Williams, 2006), vocational interest (Tay, Drasgow, Rounds, & Williams, 2009), and attitude (Carter & Dalal, 2010). Although several ideal point models have been developed (Cooper & Nakanishi, 1983; Kamakura & Srivastava, 1986; Zinnes & Griggs, 1974), the Generalized Graded Unfolding Model (Roberts, Donoghue, & Laughlin, 2000)(GGUM; Roberts, Donoghue, & Laughlin, 2000) has become popular in recent years, and is unique in that it can be applied to both dichotomous and polytomous response data.

Failure to recognize the correct item response process is detrimental to both measurement accuracy and substantive conclusions. For example, a misidentified item response process will lead to inaccurate estimation of curvilinear relationships (e.g., Carter et al., 2014; Carter, Dalal,

[1] The 2PL was invented to handle tests scored as "correct" v. "incorrect," whereas the GRM was designed to handle essay items graded by a single rater.

Guan, LoPilato, & Withrow, 2017). Applying an inappropriate measurement model (i.e., the dominance model) to ideal point response data could allow a spurious additional dimension to emerge (Davison, 1977). In addition, higher reliability can be achieved when the test is built using the appropriate response process in personality measurement (Chernyshenko, Stark, Drasgow, & Roberts, 2007). Given the fundamental differences in dominance and ideal point response processes, it is imperative to identify the correct item response process of the data before proceeding with any statistical analyses.

Tay, Ali, Drasgow, and Williams (2011) have suggested using the test-level adjusted χ^2/df ratios to identify the correct item response process. The adjusted χ^2/df ratios are extended from the ordinary χ^2 value, which measures the difference between the observed and expected response patterns within a designated IRT model. Drasgow and colleagues (1995) further developed adjusted χ^2/df ratios to calculate IRT-based model-data fit from the Pearson χ^2:

$$\chi_i^2 = \sum_k \frac{[O_{ik} - E_{ik}]^2}{E_{ik}},$$

where O_{ik} represents the observed frequency for response option k to this item, which is simply determined by counting the number of times that individuals selected this option in the sample. $E_{ik}$is the expected frequency, which can be calculated from the item response function for option *k* integrated over a standard normal density function - $f(\theta)$- multiplied by the sample size, *N*.

$$E_{ik} = N \int P(U_i = k|\theta) f(\theta) d\theta,$$

As Van den Wollenberg (1982) pointed out that χ^2 "singlet" is not sensitive to certain types of misfit (van den Wollenberg, 1982), therefore, the χ^2 "doublet" was introduced and can be computed using a two-way contingency table. For a pair of item *i* and *i'*, observed frequencies

represent response option k in two items; expected frequency of χ^2 doublet can be calculated as an extension of the χ^2 singlet:

$$E_{ii'kk'} = N \int P(U_i = k|\theta)P(U_{i'} = k'|\theta)f(\theta)d\theta,$$

Furthermore, χ^2 "triplet" was developed and can be computed using a multiway contingency table, and the expected frequency can be found as:

$$E_{ii'i''kk'k''} = N \int P(U_i = k|\theta)P(U_{i'} = k'|\theta)P(U_{i''} = k''|\theta)f(\theta)d\theta.$$

Because expected value of a noncentral χ^2 value is equal to its *df* plus N times its noncentrality parameter δ, one problem with the χ^2 value is that the value is sensitive to sample size, that is:

$$E(\chi^2) = df + N\delta.$$

To resolve this issue, Drasgow and colleagues suggested to adjust the χ^2 value to a fixed sample size of 3,000 to facilitate the comparability of these indices across different sample sizes. Therefore, the nocentrality parameter can be represented as:

$$\hat{\delta} = (\chi^2 - df)/N.$$

Finally, an adjusted $\chi/^2$ratio to fix degree of freedom and sample size of 3,000 can be calculated using the equation:

$$\text{Adjusted } \chi^2_{3000}/df = \left[df + \frac{3000(\chi^2 - df)}{N}\right]/df.$$

Since this development, adjusted χ^2/df ratios have been frequently employed in IRT-related applications to measure model-data fit (Carter & Dalal, 2010; Cho, Drasgow, & Cao, 2015a; Stark et al., 2006) (e.g., Stark, 2001; Stark, Chernyshenko, Drasgow, & Williams, 2006; Carter & Dalal, 2010; Cho, Drasgow, & Cao, 2015).

In two Monto Carlo simulation studies, Tay et al. (2011) used adjusted χ^2/df ratios to examine the relative fit of the ideal point model (GGUM) to dominance data, as well as the dominance model (2PL and GRM) to ideal point data, based on dichotomous and polytomous items (Louis Tay, Ali, Drasgow, & Williams, 2011). These studies assumed that first, if the ideal point model could fit the dominance data well, it would indicate that the ideal point model was flexible enough to fit the dominance data; second, that two models are not statistically distinguishable if they fit all data type equally well. Using simulated dichotomous responses, Tay et al. observed that both adjusted χ^2/df ratios doubles and triples could be utilized to determine whether responses followed the ideal point assumption, but that fit statistics were less likely to correctly identify the dominance data. Unlike the mixed observations from the dichotomous data, Tay and colleagues suggested that the adjusted χ^2/df ratios were capable of correctly identifying the true model for the polytomous response data.

Results from the two studies in Tay et al. (2011) conflict, indicating that although test-level adjusted χ^2/df ratios doubles and triples can accurately capture the item response process when the underlying model was the GRM, this was not the case when the underlying model was the 2PL, especially when sample sizes were in a moderate range. Theoretically, the 2PL is simply a special case of the GRM for dichotomous data, and thus the GRM is a generalization of the 2PL to polytomous item response data (Embretson & Reise, 2000). Therefore, from a theoretical perspective, findings regarding these two IRT models should be consistent.

One possible explanation for the inconsistent findings in Tay et al. (2011) may involve issues encountered by the authors regarding non-convergence of the GGUM. Table 1 displays the non-convergence rates (i.e., the number of non-converged replications in each condition / the number of total replications in each condition) by condition when fitting the GGUM to data

arising from ideal point and dominance response processes. A considerable number of replications were unsuccessful in obtaining convergence in polytomous data when the GGUM was fit to the simulated GRM data across two test lengths (i.e., 5-item and 10-item scales) and four levels of sample sizes (i.e., 500, 1,000, 1,500, and 2,000), the non-convergence rates ranged from 22.5% to 29% and 6.5% to 16% for 5-item and 10-item scales, respectively. In this experiment, Tay et al. (2011) utilized 200 expectation-maximization (E-M) outer-cycles and 30 inner-cycles in their marginal maximum likelihood (MML) estimation, and the convergence criterion was set to 0.0001 for the GGUM. Technically, parameters for a given item are updated iteratively until there is either a minor change in parameters from one iteration to the next, or the maximum limit of iterations has been reached (Roberts, Fang, Cui, & Wang, 2004). Non-convergence implies that parameter estimates do not represent a stable solution after the set number of iterations, making it difficult to draw precise conclusions from these studies. The better fit of the "correct" model is confounded with non-convergence; thus, an alternative interpretation of these results is that adjusted χ^2/df ratios doubles and triples can identify a model that did converge from a model that did not, which would be an unsurprising result that is not meaningful with regard to identifying the correct response process.

The first purpose of the current study is to re-evaluate the role of adjusted χ^2/df ratios in identifying item response process. Experiments from Tay et al. (2011) will be replicated, but with an increased number of iterations to determine the true usefulness of adjusted χ^2/df ratios. In addition to re-investigating the effectiveness of adjusted χ^2/df ratios, other common statistical indices will be evaluated, given the demand for accessible methods of identifying underlying item response processes in self-report data. The current study focuses on model-data fit statistics that are theoretically built for model comparison, and are commonly available in software,

including the 'mirt' package in R (Chalmers, 2012), IRTPro (Cai, Thissen, & du Toit, 2011), FlexMIRT (Cai, 2017), and GGUM2004 (Roberts, Fang, Cui, & Wang, 2006). Among all model-data fit statistics, the following global relative fit statistics: (a) the log-likelihood test (LL; (Wilks, 1938)), (b) Akaike information criterion (AIC; (Kullback & Leibler, 1951)), (c) Bayesian information criterion (BIC; Schwarz, 1978), (d) M2 statistics, e) RMSEA, f) Tucker-Lewis Index (TLI;), and g) Comparative Fit Index (CFI;) are identified as appropriate statistics to evaluate relative model-data fit based on their broad applications and the information-theoretical approach in model evaluation (Posada & Buckley, 2004)(e.g., Brunham & Anderson, 2002; Posada & Buckley, 2004). The major exception to this rule is the adjusted χ^2/df ratio, which is a more straightforward approach but currently can only be implemented using the MODFIT (Stark, 2001) Excel Macro, which is – as of now – outdated and does not function well on most machines. Therefore, R code is presented for easy implementation of this method of fit assessment.

In the following sections, statistical and empirical differences between dominance and ideal point response processes, as well as consequences of misidentifying the item response process, are presented. Next, the identified model-data fit statistics (i.e., LL, AIC, BIC, M2, RMSEA, TLI, and CFI) are explained. Finally, two Monto Carlo simulations are proposed. In the first study, Tay et al. (2011) study will be replicated with an increased number of iterations to re-investigate the benefits of using adjusted χ^2/df ratios in item response process identification using only fully converged models' item parameter estimates. In the second study, empirical item parameters will be used to generalize findings from Study 1. In both studies, LL, AIC, BIC, M2, RMSEA, TLI, and CFI will be obtained and evaluated in addition to the adjusted χ^2/df ratios.

CHAPTER 2

2 REVIEW OF LITERATURE

2.1 How the Response Process Makes a Difference

2.1.1 Dominance Model

Most Likert-type scales are constructed based on classic testing theory (CTT), which takes on the dominance assumption, wherein agreement to items represents higher levels of the trait. “I always keep my room clean” is an example of a dominance-based conscientiousness item. Respondents are more likely to endorse items positively when their standing on the trait continuum is higher than the level indicated by the item, and endorse negatively when their standing on the trait continuum is lower than level indicated by the item, leading to a monotonically-increasing relationship between the trait and the level of agreement. The relationship can be represented as an item response curve (IRC). An example of the IRC that follows the dominance response process is shown in Figure 1.

In this article, the popular 2PL model (Birnbaum, 1968) and its polytomous generalization, Samejima’s GRM (1969) , are utilized as exemplar dominance-based IRT models. The probability of a person endorsing an item that is calibrated by the 2PL is calculated using the equation:

$$P\left[U_i = 1 \middle| \theta_j\right] = \frac{1}{1+\exp(-1.702\alpha_i(\theta_j - b_i))},$$

where α_i and b_i represent item discrimination and item location, respectively. As an extension of the 2PL, Samejima’s GRM (1969) has been used as one of the most widely accepted dominance-based IRT model for measuring polytomous responses. GRM is considered as an “indirect” IRT

model because computing the probability for an individual responding in a category is a two-step process. What occurs in the GRM calibration is that the item is treated as a series of dichotomies. For a graded response item with four categories, these dichotomies can be comprehended as 0 vs. 1, 2, 3; 0, 1 vs. 2, 3; and, 0, 1, 2, vs. 3. Therefore, the 2PL model can be used to estimate each dichotomy with the constrain that the slopes of these curves are equal within an item (Embretson & Reise, 2000). This formulation of the GRM refers to these curves as *boundary response functions* (BRF), each of which has a threshold akin to the item location in the 2PL. The BRF can be computed:

$$P_{i,k}^{*}\left[U_i = 1 \middle| \theta_j\right] = \frac{1}{1+\exp(-1.702\alpha_i(\theta_j - b_{ik}))}, k = 2,\ldots K.$$

where one item discrimination (α_i) parameter and three b_{ik} parameters can be estimated for an item with four response categories. Note that the BRF above is simply the 2PL function described earlier. This mathematical linkage between the 2PL and the GRM establishes that the GRM is a generalization of the 2PL to polytomous item response data. Once the BRF is computed for each dichotomy, the actual probability of endorsing a given response option *k* = 0,…3 on item *i* can be computed by subtraction resulting in the *option characteristic curves (OCC)*:

$$P\left[U_{ik} \middle| \theta_j\right] = P_{i,k}^{*}(\theta_j) - P_{i,k+1}^{*}(\theta_j), k = 1, \ldots, K.$$

2.1.2 Ideal Point Model

Thurstone (1927) first suggested the law of comparative judgment, which indicates that psychological item endorsement involves comparison of a series of stimuli on the same trait continuum (Thurstone, 1927). Later, Coombs (1964) extended this idea, suggesting an unfolding theory, which maps individuals and stimuli into a common space. In this framework, individuals will only endorse the item when it is close to their own standing on the trait continuum and reject

the item when the level indicated by the item does not match their trait levels. For example, respondents who encountered the item "I tend to do just enough work to get by" could reject the item for two reasons: if their standing on the trait continuum is too far above (i.e., respondents work extremely hard on every task), or too far below (i.e., respondents do not work enough) than the level indicated by the item. This comparison response process leads to a "bell-shaped" relationship between the trait level (θ) and the probability of endorsement. An example of this IRC can be found in Figure 2. The peak of the curve represents the trait level measured from the item. One important implication of this response theory is that moderately-worded items can be adequately modeled in ideal point models, whereas in dominance models, they would appear to be unrelated to the trait.

The GGUM (Roberts et al., 2000) is the first ideal point-based IRT model that generalizes to both dichotomous and polytomous responses, and has been popular due to its highly-accessible software program GGUM2004 (Roberts, Fang, Cui, & Wang, 2004). The probabilistic function of the GGUM for an item *i* presented to a person *j* in the polytomous case can be demonstrated as:

$$P[U_i = 1|\theta_j] = \frac{\exp(\alpha_i[(\theta_j - \delta_i) - \tau_{1i}]) + \exp(\alpha_i[(2(\theta_j - \delta_i) - \tau_{1i}])}{1 + \exp(\alpha_i[3(\theta_j - \delta_i)]) + \exp(\alpha_i[(\theta_j - \delta_i) - \tau_{1i}]) + \exp(\alpha_i[(2(\theta_j - \delta_i) - \tau_{1i}])},$$

where θ_j denotes the location of respondent *j* on the latent dimension underlying responses; δ_i represents the location of item *i* on the latent continuum, and can be interpreted as the level of θ corresponding to the highest likelihood of affirming the item; α_i refers to the discrimination parameter for item *i*, and τ_{i1} indicates the location of the subjective response category threshold on the latent continuum. Probability of endorsing each option in one item can be obtained based on the item location parameter, discrimination parameter and the location of the subjective response category thresholds.

2.2 Importance of Applying the Correct Item Response Process

Since the development of the GGUM, comparisons between the dominance and ideal point models have become vital. Stark and his colleagues (2006) were among the first to investigate individuals' response process to personality items (Stark et al., 2006). They compared fit statistics of two dominance IRT models (2PL and Levin's MFSM with dominance constraints) with two ideal point models (GGUM and Levine's MFSM with ideal point constraints) to data from the Sixteen Personality Factor Questionnaire (16PF). They observed better model-data fit statistics when fitting personality items to the ideal point models, indicating that individuals adopted the ideal point response process rather than the dominance response process. Chernyshenko and colleagues (2007) later developed a measure of the conscientiousness facet of order according to ideal point principles and compared it to scales developed from the same item pool using dominance assumptions (Chernyshenko et al., 2007). This study also demonstrated that the ideal point response process (GGUM) was a more appropriate approach to interpreting personality variables than the dominance approach (2PL), resulting in higher test reliabilities across the trait continuum.

In addition to personality items, the ideal point response process is also desirable for measuring other psychological constructs reflecting typical behaviors, such as vocational interests, attitudes, and values. For example, Tay, Drasgow, Rounds, and Williams (2009) explored the item response process underlying three vocational interest inventories (Louis Tay et al., 2009). They obtained the fit of both dominance (2PL and 3PL) and ideal point (GGUM) IRT models and found that the ideal point model fit better than the dominance model in most inventories; they further observed that test scores calculated by the dominance models seemed illogical where mostly realistic or mostly social items endorsed by individuals were given similar

scores, whereas the ideal point model provided better descriptions of item responses. Similarly, Carter and Dalal (2010) fit the ideal point model (GGUM) to response data of the Work Scale of the Job Descriptive Index; they found that the ideal point response process better described work satisfaction than did the dominance response process (Carter & Dalal, 2010). Finally, Ling and colleagues (2016) investigated the item response process to scales measuring values (Ling et al., 2016). They compared the fit of dominance (Generalized partial credit model [GPCM]) with ideal point (GGUM) IRT models to data from the Circumplex Scales of Interpersonal Values (CSIV). Results indicated that the ideal point approach best described the response process to value items.

In contrast with the ideal point response process, the dominance response process is more appropriate when measurement reflects maximal performance. For example, Cho, Drasgow, and Cao (2015) investigated three measures of emotional intelligence (EI): the Wong and Law Emotional Intelligence Scale (WLEIS), the Schutte Self-Report Emotional Intelligence Test (SEIT), and the Trait Emotional Intelligence Questionnaire (TEIQue) (Cho et al., 2015). They fit both dominance (GRM) and ideal point (GGUM) IRT models to these measures. Cho et al. demonstrated that the dominance model (GRM) showed better fit to the WLEIS and fit better with most subfacets of the SEIT than the ideal point model (GGUM).

Given the fundamental differences between dominance and ideal point models, researchers have argued that misidentifying the response process affects statistical outcomes. For example, using emotion-related (e.g., happy-sad) responses, Tay and Drasgow (2012) showed how misapplications of principal component analysis (PCA) to data that originates from an ideal point response process could lead to inaccurate statistical results (Tay & Drasgow, 2012). Using Monto Carlo simulation techniques, they demonstrated that applying an inappropriate

measurement model (dominance model) to ideal point response data could lead to the emergence of a spurious additional dimension, echoing earlier theory on the topic (Davison, 1977; Maij-de Meij, Kelderman, & van der Flier, 2008) (see Davison, 1977; Maij de Meij, Kelderman, & van der Flier, 2008).

Additionally, response process misidentification influences substantive conclusions. Researchers long believed personality-performance relationship to be linear (Ones, Dilchert, Viswesveran, & Judge, 2007). However, mixed research findings suggest that personality-performance relationship could be curvilinear. For example, some researchers have argued that too much conscientiousness may not be a good thing (Grant & Schwartz, 2011; Le et al., 2011). Recently, Carter and colleagues (2014) demonstrated that incorrect assumption of item response process usage can affect the conscientiousness-job performance relationship (Carter et al., 2014). Across two studies, the curvilinear relationship was revealed 100% of the time when the ideal point-based assumption (GGUM) was employed to score conscientiousness items, whereas mixed results were found when scoring based on the dominance assumption (GRM), reflecting the 50-50 split suggested by existing literature on curvilinear conscientiousness-performance relationships. They further suggested that selection through predicted performance using the ideal point-based response process leads to more favorable hiring outcomes. Additionally, Carter and colleagues (2016) tested curvilinear relationships between conscientiousness and psychological well-beings. They found that individuals with high levels of conscientiousness that overlap with obsessive-compulsive tendencies (e.g., punctiliousness, ruminative deliberation) had lower well-being (e.g., life satisfaction, self-esteem). Mixed results were obtained when the dominance-based scoring method was applied to score conscientiousness items, whereas more curvilinear relationships were detected using the ideal point response process.

A firm explication of the pitfalls of misidentifying the item response process was achieved in a series of Monte Carlo simulation studies by Carter et al. (2017) (Carter et al., 2017). Specifically, they stated that a model-appropriate scoring approach was essential for testing curvilinearity. In the first study, they simulated responses based on the ideal point model (GGUM), varying the degree of items (Strong Dominance, Weak Dominance, Weak Ideal Point, and Strong Ideal Point conditions). The accuracy of curvilinear regression results for predictors whose responses were generated by the ideal point models. Three scoring methods were compared: sum-score, GRM, and GGUM. Type I error rates for both GGUM and GRM were very close to 5% across all item locations. Power for testing curvilinearity of the ideal point-based scoring (GGUM) was high across all item locations, whereas the dominance-based scoring (sum-score and GRM) showed decreased power as the item location moved from dominance to the ideal point. With the inspection of bias and *RMSE* for curvilinear regression parameters, they found that estimation results were highly accurate across simulation conditions when the appropriate model (GGUM) was selected for scoring. Although the GRM showed similar accuracy to the GGUM when item locations were extreme (in the Weak and Strong Dominance conditions), accuracy dropped drastically when item locations distributed more uniformly.

In the second study, Carter et al. simulated responses arising from the dominance model (GRM) with the goal of investigating how well each scoring method (sum-score, GRM, and GGUM) would be able to capture curvilinearity. The accuracy of curvilinear regression results for predictors whose responses were generated by the dominance model. Type I error rates for the GRM were at 5% across conditions. All three scoring methods showed high power—ranging from 93% to 97%—for testing curvilinearity. Although both the GRM and GGUM showed unbiased estimates of curvilinear regression parameters, higher estimation error for the GGUM

was observed because the designated response model was based on the dominance approach. According to findings from these two simulations, Carter and colleagues demonstrated the importance of identifying and applying appropriate scoring techniques that best describe the response process.

Based on the research summarized here, it is vital for researchers and practitioners to correctly identify response process in self-report data before proceeding to further statistical investigations. Failure to understand the underlying item response process can impact statistical outcomes and substantive conclusions. Despite the demand for accessible methods of identifying underlying item response processes of data, at present, guidance on item response process identification is absent from the literature. In the following section, identified model-data fit statistics (e.., LL, AIC, BIC, M_2 statistics, RMSEA, TLI, and CFI) are reviewed, and will be further examined in simulation studies to examine their effectiveness in assisting with item response process identification.

2.3 Model-Data Fit Statistics

Given the importance of identifying response process in self-report data, assessment of model-data fit of a set of responses is always a challenging task in IRT models because accessible options are very limited in commercial software. Because LL, AIC, BIC, M_2 statistics, RMSEA, TLI, and CFI have been selected as appropriate statistics to assess model-data fit in addition to the adjusted χ^2/df ratios based on their broad applications and the information-theoretical approach in model evaluation (Posada & Buckley, 2004)(e.g., Brunham & Anderson, 2002; Posada & Buckley, 2004). Therefore, a comprehensive review focuses on these statistics is presented below.

Log-Likelihood (LL). The likelihood function is a function of the parameters of a statistical model, given specific observed data. For many applications, the natural logarithm of the likelihood function, called the log-likelihood (LL), is more convenient to work with. This is because natural logarithm is a monotonically-increasing function by nature. The log function achieves its maximum value at the same point as the function itself, and therefore, the LL can be used in place of the MLE and related techniques. In statistics, LL is often used to test whether a simplifying assumption for a model is valid, as when two or more parameters are assumed to be related. LL is built upon the maximum likelihood estimation (MLE), which maximizes the probability given the defined parameter estimates. For example,

$$L\left(x_1, x_{2,} \ldots x_n \middle| \theta\right) = P_{x_1 \ldots x_n}\left(x_1, x_{2,} \ldots x_n \middle| \theta\right),$$

where $x_1, x_{2,} \ldots x_n$ represent random samples from a distribution with a parameter θ. For the model comparison purposes, LL is only meaningful when comparing LL values obtained across different models. The higher LL indicates a better model.

Akaike Information Criteria (AIC). In 1951, Kullback and Leibler (K-L) first attempted to quantify the meaning of "information" where they discussed to minimize the loss of information for model selection. K-L information was soon proposed to represent the information loss when approximating the reality given data. In practice, K-L information intends to measure informational distance between two probability distributions – a model distribution and a distribution based on the data. The K-L information can be represented by the equation:

$$I(p; q) = \sum_{i=1}^{n} pi(logp_i - logq_i),$$

where discrete distributions given by probability vectors p (p_1,…p_n) and q(q_1,…q_n); log is the natural logarithm.

Later in 1973, Akaike found a formal relationship between the K-L information and the likelihood theory (see deLeeuw, 1992), and a concept of incorporating MLE with the K-L information to assist model selection was introduced. Ultimately, Akaike developed an information criterion to estimate the K-L information based on the log-likelihood, the AIC can be calculated as:

$$AIC_i = -2(logL_i) + 2p_i,$$

where *i* represents each model, p_i is the number of parameters included in the model *i*, $2p_i$ represents a penalty for including additional predictors in the model. There is always a tradeoff for creating a better fit model. The $-2(logL_i)$ intends to reward the fit between the model and the given data. However, the greater amount adds to $2p_i$ when more parameters are added to the model so that the AIC value is increased.

AIC is a comparative fit index to which the fit of a model is only interpretable when comparing AIC values across several models. Assuming a set of candidate models has been defined and an AIC value can be computed for each candidate model; these candidate models can be easily ranked from best to worst based on their AIC values. AIC is a simple, but compelling concept built upon strong theoretical foundations (i.e., K-L information and likelihood theory), and therefore, the AIC is widely employed by researchers and practitioners for model selection purposes.

Bayesian Information Criterion (BIC). Although it is possible to increase the likelihood by adding more parameters when fitting models, this may lead to an overfitted model. BIC (Schwarz, 1978) intends to solve this problem by introducing a penalty term based on the number of parameters. Although the BIC is closely related to the AIC, the BIC was designed under the Bayesian context. A greater penalty term is introduced in the BIC than the AIC.

Therefore, selecting a simpler model that contains less parameters is encouraged. The BIC can be computed by the equation:

$$BIC_i = -2(logL_i) + p_i log\ n.$$

where *i* represents each model, p_i is the number of parameters included in the model *i*, and *n* represents the sample size. There is always a tradeoff for creating a better fit model. The $-2(logL_i)$ intends to reward the fit between the model and the given data. $p_i log\ n$ represents a penalty for including additional predictors in the model. The greater amount adds to this component when additional parameters are added to the model, so that the BIC value is increased. Although researchers argue the favorability of using the AIC over the BIC (e.g., Yang, 2005), the BIC is still largely used as a popular model selection index. As with the AIC, the BIC is a comparative fit index as well. Therefore, a BIC value can be computed for each candidate model to facilitate model selection.

M_2 Statistics. Recently, limited-information goodness-of-fit testing has received increased attention in the psychometrics literature. In contrast to full-information test statistics such as Pearson χ^2 test statistic or the likelihood ratio test statistic G^2, these limited-information tests utilize lower-order marginal tables rather than the full contingency table. When the number of items is large, or the number of respondents is small, the contingency table for possible responses is sparse. Therefore, χ^2 and G^2 cannot be trusted to test for the lack of fit (e.g., Maydeu-Olivares & Joe, 2005). The limited-information test statistics (Reiser, 1996; Reiser & Lin, 1999) is more practical because they only use low-order marginal information in the contingency table to evaluate the model-data fit (Cai & Hansen, 2013). For example, as Liu, Tian, and Xin (2016) presented, the lower-order marginal probabilities of a three-item test can be shown as:

$$\dot{\pi} = \begin{pmatrix} \dot{\pi_1} \\ \dot{\pi_2} \\ \dot{\pi_3} \\ \dot{\pi_{1,2}} \\ \dot{\pi_{1,3}} \\ \dot{\pi_{2,3}} \\ \dot{\pi_{1,2,3}} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{(0,0,0)} \\ \pi_{(1,0,0)} \\ \pi_{(0,1,0)} \\ \pi_{(0,0,1)} \\ \pi_{(1,1,0)} \\ \pi_{(1,0,1)} \\ \pi_{(0,1,1)} \\ \pi_{(1,1,1)} \end{pmatrix}$$

The limited-information goodness-of-fit test statistics compare the model-predicted marginal probabilities and the corresponding sample-observed counterparts. M_2 statistics has been used as one popular limited-information test statistics, and it can be calculated as shown below (Maydeu-Olivares & Joe, 2005):

$$M_2 = N(p_2 - \widehat{\pi_2})'\widehat{C_2}(p_2 - \widehat{\pi_2})$$

where π is the vector of cell probabilities, $\widehat{\pi_2}$ is the maximum likelihood estimator of π, p is the vector of sample cell proportions, and $\widehat{C_2} = \widehat{\Delta_2^{(c)}} * 1/(\widehat{\Delta_2^{(c)}}' \hat{r}_2 \widehat{\Delta_2^{(c)}}) * \widehat{\Delta_2^{(c)}}'$. $\widehat{\Delta_2^{(c)}}$ is an $R * (R - F)$ orthogonal complement to $\widehat{\Delta_2}$ (Browne, 1984), the corresponding *df* can be calculated as *R- F* (Khatri, 1966)

Root Mean Square Error of Approximation (RMSEA). This absolute measure of fit is based on the non-centrality parameter. The RMSEA can be found as:

$$RMSEA = \sqrt{\frac{\chi^2 - df}{df(N-1)}}$$

where N represents the sample size and *df* represents the degrees of freedom of the model. Its penalty for complexity is the χ^2 to *df* ratio. MacCallum, Browne and Sugara (1996) have suggested to use .01, .05, and .08 to indicate excellent, good and average fit, respectively. However, others have suggested to use .10 to be the cutoff for poor fitting models. There is greater sampling error for models with small *df* and low N, especially for the former. Therefore,

models with small *df* and low N can have artificially large values of the RMSEA (Kenny, 2015). For instance, a chi square of 2.098 (a value that is not statistically significant), with a *df* of 1 can yield an RMSEA of .126. Due to this reason, Kenny, Kaniskan and McCoach (2014) suggest to not even compute the RMSEA for models with the low *df*. In the current study, the RMSEA was not computed when the *df* was considered low.

Additionally, use of confidence intervals can help understand the sampling error in the RMSEA. A confidence interval can be computed for the RMSEA. The width of the confidence interval is very informative about the precision in the estimate of the RMSEA. Ideally the lower value (i.e., less than .05) of the 90% confidence interval (less than .05) and the upper value is not very large, i.e., less than .08. Nevertheless, because theses RMSEA cutoffs were developed based on the Structural Equation Modeling (SEM), they do not necessarily apply to the IRT models.

Tucker-Lewis Index (TLI). Originating from Tucker and Lewis (1973), Bentler and Bonett (1980) applied TLI (also known as the non-normed fit index or NNFI) to covariance structure analysis and claimed that it can be used to compare a model across samples. TLI is calculated as:

$$TLI = \frac{\frac{\chi^2}{df}(Null\ Model) - \frac{\chi^2}{df}(Proposed\ Model)}{\frac{\chi^2}{df}(Null\ Model) - 1},$$

which depends on the average size of the correlations in the data. The TLI will not be very high if the average correlation between variables is not high. TLI can exceed the 0 to 1 range. However, if the index is greater than 1, it is set at 1. Anderson and Gerbing (1984) show that TLI values close to 1 indicates a correctly specified model, but in small samples (sample size smaller than 100) its value is underestimated (that is, indicates a bad fit for an acceptable model) and has

large sampling variability. Hu and Bentler (1999) recommend a cutoff value of TLI close to .95. When it comes to the model comparison, the higher the TLI indicates a better model (Hu & Bentler, 1995). However, theses CFI cutoffs were developed based on the Structural Equation Modeling (SEM), so they do not necessarily apply to the IRT models.

Comparative Fit Index (CFI). To avoid TLI's problems concerning underestimation of fit and considerable sampling variability in small samples, Bentler proposed the CFI in 1988. In addition, Bentler (1990) conducted a simulation study to compare TLI, CFI, the Normed Fit Index (NFI) and the Incremental Fit Index (IFI) and concluded that CFI was the best index. TLI indicates more misspecification, and CFI has the advantages of having a 0 to1 range and smaller sampling variability. CFI is defined as:

$$TLI = \frac{(\chi^2 - df)(Null\ Model) - (\chi^2 - df)(Proposed\ Model)}{(\chi^2 - df)(Null\ Model)}$$

Hu and Bentler (1999) recommend a cutoff value of CFI close to .95. Both TLI and CFI are incremental fit indices, which measure the improvement of fit by comparing a H_0 model with a more restricted baseline model (Hu & Bentler, 1999). When it comes to the model comparison, the higher the TLI indicates a better model (Hu & Bentler, 1995). However, theses TLI cutoffs were developed based on the Structural Equation Modeling (SEM), so they do not necessarily apply to the IRT models.

Theoretically, LL, AIC, BIC, M_2 statistics, RMSEA, RMSEA overlap, TLI, and CFI are possible to identify the correct item response process. Both AIC and BIC are log-likelihood-based statistics. LL are comparative fit statistics to which a higher value indicates a better model-data fit. AIC and BIC are comparative fit statistics as well to which a lower value indicates a better model-data fit. M_2 statistics, RMSEA, TLI, and CFI can select a better model when the values are smaller. In two proposed Monto Carlo simulation studies, simulated conditions were

implemented to evaluate the effectiveness of LL, AIC, BIC, M_2 statistics, RMSEA, TLI, CFI as well as the adjusted χ^2/df ratios in identifying the correct response process of the self-report data. The study assumption is that if any of these statistics consistently points to the true response model across simulation conditions, it could be concluded that the index is capable to identify the correct response process. However, two models are not statistically distinguishable if the statistics fit all data type equally well.

CHAPTER 3

3 METHODS

3.1 Study 1 Method

The first goal of the current study is to examine the capabilities of LL, AIC, BIC, M_2 statistics, RMSEA, TLI, and CFI for identifying the correct response process model of the data by comparing relative model-data fit statistics. In addition, the second goal of the current study is to re-evaluate the capabilities of singlets, doublets and triplets χ^2/df ratios when the non-convergence issues observed in the Tay et al. (2011) study are resolved. This study replicated the study design of the Tay et al., and all aforementioned model-data fit statistics were evaluated based on both dichotomous and polytomous response data arising from dominance and ideal point response processes. In order to resolve the non-convergence issues observed by Tay et al., this study was conducted with an increased number of iterations to ensure the model convergence. Across all simulation conditions, the convergence criterion was set to .0001; the number of maximum iterations was specified as 1,000, 000. All replications converged in the current study.

3.1.1 Dichotomous IRT Models

3.1.1.1 Simulation design

As in the Tay et al. (2011) study, a factorial 4 (Sample Size: 250, 500, 1000, 2000) × 2 (Scale Length: 15, 30) × 2 (Sample Type: Calibration Sample and Cross-Validation Sample) was used, for a total of 16 conditions. Within each condition, 100 calibration samples were used to simulate 100 calibration datasets and 100 cross-validation response datasets.

3.1.1.2 Item parameter and data generation

Parameters of dichotomously scored items were generated based on the 2PL to accommodate the dominance response process. The item discrimination parameter, α_i, was created by randomly sampling a value from a log-normal (0, 0.5) distribution and dividing by 1.702, which is the scaling parameter. The item location parameter, $b_{i,\,was}$ simulated from a uniform (-2, 2) distribution.

Item parameter generation of the ideal point response process closely followed the GGUM procedures proposed by Roberts, Donoghue, and Laughlin (2002), and item discrimination parameter, $\alpha_{i,}$ was sampled from a uniform (0.5, 2) distribution. Subjective threshold parameter, τ_i was simulated from a uniform (-1.4, 0.4) distribution. Finally, the location parameter, δ_i , was sampled from -2 to 2 in equal increments with the four middle values removed to accommodate the fit of the 2PL (see Tay et al., 2011). They had to be excluded because they could not be estimated using the dominance IRT models.

Dichotomous response data was generated by drawing θ values from a standard normal (0, 1) distribution. Item response probability for a given IRT model was compared with a random uniform (0, 1) number *r*. A positive item response (1) was simulated when response probability was larger than *r*, and a negative response (0) occurred when response probability was less than *r* (Harwell, Stone, Hsu, & Kirisci, 1996).

3.1.2 Polytomous IRT Models

3.1.2.1 Simulation design

A study of factorial 4 (Sample Size: 500, 1000, 1500, 2000) × 4 (Scale Length: 5, 10, 15, 20) × 2 (Sample Type: Calibration Sample and Cross-Validation Sample) was designed that contained 32 conditions. Within each condition, 100 calibration samples were used to simulate 100 calibration datasets and 100 cross-validation response datasets.

3.1.2.2 Item parameter and data generation

Item parameters underlying the dominance response model were simulated based on the GRM. For each item, the item discrimination parameter, α_i, was generated from a uniform (0.5, 0.8) distribution, and four b_i parameters were sampled from uniform distributions of (-2.0, -1.0), (-1.0, 0.0), (0.0, 1.0), and (1.0, 2.0). These values were selected based on a previous study simulating personality data under the GRM (see Reise & Waller, 1990).

Item parameter generation of the GGUM closely followed the procedures proposed by Roberts, Donoghue, and Laughlin (2002). For each item, α_i was sampled from a uniform (0.5, 2) distribution, and τ_{iC} was generated from a uniform (-1.4, 0.4) distribution. True τ_{ik} parameters were then generated using a recursive equation:

$$\tau_{ik-1} = \tau_{ik} - .25 + e_{ik-1}, \qquad \text{for } k = 2, 3, \ldots C,$$

where e_{ik-1} denotes a random error term generated from a N (0, .04) distribution, and δ_i were sampled from -2 to 2 in equal increments with the four middle values removed. Finally, response data was simulated by comparing a random uniform number (0, 1) with a cumulative distribution (see Harwell, Stone, Hsu, & Kirisci, 1996).

3.1.3 Fitting IRT models

Both the 2PL (dichotomous), GRM (polyltomous), and GGUM (dichotomous and polytomous) were fit to the generated data using the 'mirt' package developed in R (Chalmers, 2012). The 'mirt' package can handle both dichotomous and polytomous responses. To ensure the model convergence and the comparability of item parameter estimations, the convergence criterion was set to .0001; the number of maximum iterations was specified as 1,000, 000. All replications were converged in the current study. Item parameters were obtained along with the model-data fit statistics, including the LL, AIC, BIC, M_2 statistics, RMSEA, TLI, and CFI.

In addition, an R (R Development Core Team, 2012) program was written to calculate test-level single, double and triple adjusted χ^2/df ratios using the item parameters retrieved from the 'mirt' package. The R code for calculating GGUM-based adjusted χ^2/df ratios can be found in Appendix A. Higher values of LL, TLI and CFI indicate better model-data fit, while lower values of AIC, BIC, M_2 statistics, RMSEA, and adjusted χ^2/df ratios indicate better model-data fit.

3.2 Study 2 Method

The first goal of this study was to replicate the Tay et al. study using exactly the same study setting. I hope to compare the results from this study with those of Study 1 to examine two issues observed by Tay et al.: a) a large number of non-convergences were observed when the estimation was based on the polytomous responses, but almost no non-converged replications were found when the items were dichotomously scored; and b) test-level doublets and triplets adjusted χ^2/df ratios could identify the true model accurately when the items were polytomously scored, but the adjusted χ^2/df ratios did not work well when using the dichotomous data. Because the GRM was a generalization of the 2PL, theoretically, the findings were expected to be consistent across two response types. The inconsistent observation of the adjusted χ^2/df ratios

may have been due to the fact that non-convergence was observed in the polytomous data but not in the dichotomous data. The second goal of the current study was to examine the change in the results in Study 1 when the current study results were contaminated with non-converged parameter estimates. The data simulated for Study 1 were used in Study 2 to ensure a highly controlled study. Nevertheless, in the current study, the number of iterations was specified as 200, and the convergence criterion was set to .0001 to replicate Tay et al. Subsequently, both dominance and ideal point IRT models were fit to the simulated dichotomous and polytomous data. The LL, AIC, BIC, M_2 statistics, *df*, RMSEA, TLI, and CFI were calculated using the 'mirt' package in R, and the test-level singlets, doublets and triplets adjusted χ^2/df ratios were calculated by the same R program implemented in Study 1.

CHAPTER 4

4 RESULTS

4.1 Study 1 Results

4.1.1 Dichotomous IRT Models

Table 2 and Table 3 display the averages of the LL, AIC, BIC, M_2 statistics, *df*, RMSEA, RMSEA overlap, TLI, CFI, test-level singlets, doublets, and triplets adjusted χ^2/df ratios across replications when the true model was the 2PL and the GGUM. Results can be found in Table 2 for both 2PL and GGUM fit to the responses that arose from the dominance model (i.e., the 2PL model). Across all simulation conditions, fitting the 2PL to the dominance-based dichotomous responses resulted in average LL of -11863.50, AIC of 23817.00, BIC of 24022.25, M_2 statistic of 248.89, RMSEA of .01 (90% CI = .00-.02), TLI of 1.00, and CFI of .99. Finally, the means of singlets, doublets and triplets χ^2/df ratios were 2.92, 1.90, and 1.62, respectively. In contrast, these model fit statistics changed when fitting the GGUM to the dominance response data. The average LL changed to -11886.86, AIC increased to 23908.72, BIC increased to 24216.60, M_2 statistic increased to 248.89, and RMSEA dropped to .00 (90% CI = .00-.02). Both averages TLIs of CFIs were 1.00. Higher means of singlets, doublets and triplets χ^2/df ratios were observed; they increased to 3.79, 2.29, and 1.85, respectively.

Table 3 displays the results from fitting both 2PL and GGUM to the data that arose from the ideal point response process (i.e., the GGUM model). When fitting the 2PL to the ideal point data, the average LL was -13296.64, AIC was 26683.29, BIC was 26888.54, and M_2 statistic was 931.00. The mean RMSEA was .05 (90% CI = 0.04 – 0.06), TLI was .80, and CFI was .82.

Means of singlets, doublets and triplets χ^2/df ratios were 11.94, 6.04, and 4.80, respectively. However, when the GGUM was implemented as both the true model and the calibrated model, the results changed drastically. The average LL increased to -13113.57, while the average AIC, BIC, M_2 statistic, and RMSEA decreased to 26362.13, 26670.01, 262.82, and .01, respectively. Both means of TLI and CFI increased to .99. Means of singlets, doublets and triplets χ^2/df ratios dropped to 2.51, 1.68, and 1.43, respectively.

Table 4 summarizes the results from Table 2 and Table 3, illustrating how many times each of these model-data fit statistics pointed to the correct model across all replications in each condition. When the true model was the 2PL, the LL identified the true model for 52% of the replications correctly. Additionally, the LL could identify more replications correctly in larger samples. Both AIC and BIC pointed to the correct response process almost 100% of the replications regardless of the sample size and the scale length. The RMSEA, RMSEA overlap, TLI, and CFI misidentified almost 50% of the replications. The M_2 statistic and adjusted singlets, doublets and triplets χ^2/df ratios misidentified the response processes in almost all replications. Across all calibration and validation samples, M_2 statistic only pointed to the correct model 1.81% of all replications, which was a strikingly low performance. Adjusted singlets, doublets and triplets χ^2/df ratios identified the model correctly 22.94%, 19.81%, and 16.81% of all replications, respectively.

In contrast, the performance of these model-data fit statistics changed drastically when the true model was the ideal point response process. The M_2 statistic, RMSEA, TLI, CFI, adjusted singlets, doublets and triplets χ^2/df ratios could identify the true model with nearly 100% accuracy. The LL, AIC and BIC recognized more true models in larger samples (i.e., sample sizes of 1000 and 2000). Finally, RMSEA overlap failed to differentiate the true model.

4.1.2 Polytomous IRT Models

Both Table 5 and Table 6 present the means of LL, AIC, BIC, M_2 statistics, *df*, RMSEA, RMSEA overlap, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the dominance model (i.e., GRM). Table 5 contains the results when the calibrated model was the GRM. Due to the insufficient sample size and scale length, the M_2 statistics, *df*, RMSEA, TLI, and CFI were not reported. Across all simulation conditions, the average LL was -23726.3, AIC was 47577.55, BIC was 47890.62, M_2 statistic was 77.92, the RMSEA was .01 (90% CI = .00 – .02), TLI was 1.00, and CFI was .97. Means of singlets, doublets and triplets χ^2/df ratios were 3.84, 2.96, and 2.75, respectively. In addition, Table 6 reports the results when the calibrated model was the GGUM. The average LL decreased to -23825.90. The mean AIC and BIC increased to 47801.87 and 48177.55. Meanwhile, the mean M_2 statistic fell to 62.71, and the mean RMSEA remained at .01 (90% CI = .00 – .02). Average TLI and CFI both decreased to .89. Finally, the averages of singlets, doublets and triplets χ^2/df ratios rose to 4.67, 3.65, and 3.43, respectively.

Summarizing the results presented in Table 5 and Table 6, Table 7 shows how many times each of these model-data fit statistics pointed to the correct model across all replications in each condition. As when using the dichotomous data, the AIC and BIC outperformed other fit statistics, pointing to the correct model almost 100% of the time. When the sample size and scale length were sufficient, the LL could differentiate the response process as well as the AIC and BIC. In addition, 70.38% and 60.88% of the time, RMSEA and RMSEA overlap were able to identify the correct item response process. However, M_2 statistics, TLI, CFI and all χ^2/df ratios correctly identified the item response process in less than 50% of the replications.

Additionally, Table 8 and Table 9 present the means of LL, AIC, BIC, M_2 statistics, *df*, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the ideal point response process (i.e., GGUM). Table 8 displays the results when the calibrated model was the GGUM. The average LL was -20454.00, average AIC was 41058.01, mean BIC was 41433.69, average M_2 statistic was 784.87, average RMSEA was .08 (90% CI = .07 – .09), average TLI was 0.93, and the mean CFI was .95. Means of singlets, doublets and triplets χ^2/df ratios were 1.65, 1.52, and 1.49, respectively. Table 9 presents the results produced when the calibrated model was the GRM. The average LL decreased to -21031.40. At the same time, the mean AIC and BIC increased to 42187.81 and 42500.88, while the mean M_2 statistic increased to 1838.51. The mean RMSEA also increased to .13 (90% CI = .12 – .13). Both the average TLI and CFI dropped to .88 and .90. Finally, the average singlet, doublet and triplet χ^2/df ratios increased to 8.63, 5.50, and 4.90, respectively.

Table 10 summarizes the findings listed in Table 8 and Table 9, indicating how many times these fit statistics pointed to the correct model. All χ^2/df ratios outperformed other model fit statistics; on average, the χ^2/df ratio could recognize the correct response process 98.39% of the time. LL, AIC and BIC performed slightly poorer than the χ^2/df ratios, correctly identifying the response process 83.75%, 83.31%, and 82.06% of the replications, respectively. Less than 80% of the time, M_2 statistics, RMSEA, TLI, and CFI were able to specify the correct response process. RMSEA overlap failed to capture the response process across nearly all simulations. Overall, these results are consistent with those from when the dichotomous response data was used.

4.2 Study 2 Results

4.2.1 Non-Convergence Ratios

Table 11 and Table 12 display the non-convergence rates (i.e., the number of non-converged replications in each condition/ the number of total replications in each condition) when fitting the dominance model and the ideal point model to the data arising from both response processes. Table 11 represents the non-convergence rates when fitting the dominance model to data arising from both response processes. Under dichotomous scoring, non-convergence replications only occurred when the sample size was relatively small, and the non-convergence rates ranged from 0% to 2%. When using the polytomous response data, the highest non-convergence rates were observed when fitting the dominance model to the GGUM data with the 5-item scale, as the non-convergence rates ranged from 1.5% to 8.5%. When more items were added to the scale, the occurrence of the non-convergence almost completely vanished. The results presented in Table 11 were consistent with the findings from Tay et al.

Table 12 displays the non-convergence rates by condition when fitting the ideal point model to data arising from both response processes. When fitting the ideal point model to the 2PL for the 15-item scale, the non-convergence rates ranged from 28.5% to 59.5% across four simulated sample sizes. For the 30-item scale, the non-convergence rates ranged from 31.5% to 70%. Similar non-convergence patterns were also observed when fitting the ideal point model to the GRM-based polytomous response data. For the 5-item scale, the non-convergence rates ranged from 23.5% to 63%. As more items were added to the scale, higher non-convergence rates were observed. For the 20-item scale, these rates ranged from 67.5% to 90.5%. The non-convergence observations were somewhat consistent with the findings identified in Tay et al., as

they observed high non-convergence in the polytomous data (i.e., ranged from 6.5% - 29%) but not in the dichotomous data (i.e., ranged from 0% - 6.5%).

On the contrary, the high non-convergence rates were not observed when fitting the ideal point model to the GGUM data across the two response types (i.e., either dichotomous or polytomous responses). Using the dichotomous response data, the non-convergence observations ranged from 0% to 2.5% of the total replications, whereas the rates ranged from 0% to 6.5% when fitting the model to the polytomous responses. These results were consistent with the findings from Tay et al. (2011).

4.2.2 Fitting the Dominance Model to the Data

Regardless of response type, because only a few non-converged replications were discovered when fitting the dominance model to the data arising from either of the two response processes, almost identical model-data fit statistics were found as observed in Study 1 when ensuring the full convergence. The results are displayed in Tables 13, 15, and 18. The highest non-convergence rates were found when fitting the dominance model to the GGUM-based polytomous responses with the 5-item scale, ranging from 0% - 8.5%. Due to small contaminations, nuanced differences were found across all model-data fit statistics when comparing the results from Study 2 with Study 1.

4.2.3 Fitting the Ideal Point Model to the Data

When fitting the ideal point model to the GGUM data, a small portion of replications failed to converge. The non-convergence rates ranged from 0% to 6.5% across all simulation conditions. Because of these slight contaminations, means of all model-data fit statistics did not change as much as in Study 1. These results are presented in Tables 13, 14, and 17 for both dichotomous and polytomous responses.

However, a large number of non-converged replications were found when fitting the ideal point model to the dominance model-based data. High non-convergence rates existed regardless of the response types, and even higher rates were detected using the polytomous data. Table 16 presents the averages of all model-data fit statistics across all simulation conditions. Given the large number of contaminations, the averages of these model-data fit statistics were slightly different from the ones observed in Study 1. Means of the LLs changed from -23825.90 to -23848.60, average of the AICs increased from 47801.87 to 47847.20, average of the BICs increased from 48177.55 to 48222.87, and average of the M_2 statistics, RMSEA, and RMSEA overlap remained almost the same across two studies. Averages of TLI and CFI both increased to .89. The test-level singlets, doublets and triplets adjusted χ^2/df ratios were decreased from 5.20, 4.10, and 3.86 to 4.67, 3.65 and 3.43, respectively.

4.2.4 Capabilities of Model-Data Fit Statistics Pointing to the Correct Model with Non-Converged Replications

Given the large number of non-converged replications observed in certain conditions, Table 19 updates the percentages of how many times the model-data fit statistics pointed to the correct model when the responses were dichotomously scored. The LL, AIC and BIC were still superior when the sample size and the scale length were sufficient. The AIC and BIC pointed to the correct model almost 100% of the time when the true model was the 2PL. M_2 statistics, RMSEA, TLI, CFI, test-level singlets, doublets, and triplets adjusted χ^2/df ratios were only able to indicate the correct model when the true model was the GGUM. The RMSEA overlap failed to recognize the true model nearly 100% of the replications.

Finally, Tables 20 and 21 present results based on the polytomous data. When the true model was the GRM, AIC and BIC detected the true model almost 100% of the replications. The

LL increased its performance when longer scale lengths were tested. The RMSEA, RMSEA overlap, TLI, CFI, and test-level singlets, doublets, triplets adjusted χ^2/df ratios could recognize the true model nearly 70% of the replications, especially when the sample was sufficiently large, but the M_2 statistics failed to differentiate the true model completely. When the true model was the GGUM, test-level singlets, doublets, triplets adjusted χ^2/df ratios identified the correct model close to 100% of the time. The LL, AIC, BIC, and M_2 statistics could differentiate the model 83.72%, 83.1%, 82.06%, and 78.69% of the replications, respectively. RMSEA, TLI, and CFI were worst at identifying the correct underlying model, ranging from 70.56% to 73.81%.

CHAPTER 5

5 DISCUSSIONS

5.1 Study 1 Discussions

The first goal of the current study was to examine the capabilities of these popular model-data fit statistics in identifying the correct response process. The main conclusion from this study is that no single universal model-data fit statistic can be used to identify the correct item response process based on the results observed from all simulation conditions. Among all model-data fit statistics examined, the LL, AIC, and BIC performed best when the sample size was sufficient to estimate the items. The results are highly consistent regardless of the response type (i.e., dichotomous and polytomous response data). When the true model was the dominance model, the LL failed to recognize the correct response process at smaller sample sizes (i.e., sample sizes of 250 and 500) with shorter scale length. However, its performance improved drastically when the sample size increased. Furthermore, AIC and BIC captured the true model almost 100% of the time regardless of sample size and scale length. When the true model was the ideal point model, the LL recognized the response process correctly 86.44% of the replications in dichotomous data and 83.75% of the time in polytomous data at any sample size and scale length. Across all variations in sample sizes and scale lengths, the AIC and BIC also outperformed other examined statistics. Nevertheless, both AIC and BIC did slightly worse in this scenario than when the true model was the dominance model.

Given their extremely poor performance in certain conditions, M_2 statistics, RMSEA, RMSEA overlap, TLI and CFI should not be relied upon to identify the underlying response

process of the data. Even though the M_2 statistics identified the most conditions correctly when the underlying response process was the ideal point, M_2 statistics failed to capture almost any response processes correctly when the true model was the dominance model. RMSEAs pointed to more correct models when the true model was the ideal point response process, especially when using the dichotomous data. When the true model was the dominance model, RMSEAs pointed to the correct model 70.38% of the time when the items were polytomously scored, and this rate dropped to 58.50% when using the dichotomous data. The RMSEA overlap failed to identify the true model nearly 100% of the time. The overall performance of the RMSEA and RMSEA overlap were not as strong as the LL, AIC and BIC.

TLI and CFI consistently had similar success identifying the true model. Both could identify more correct models than the dominance model when the true model was the ideal point. When the true model was the 2PL, the TLI and CFI identified the true model 47.15% and 50.12% correctly, and these rates remained almost at the same level (i.e., 48.63% of the TLI and 58.25% of the CFI) when using the polytomous data. Given the low performance, TLI and CFI are not appropriate methods to differentiate the underlying response process of the data. Overall, given the inconsistent findings observed in M_2 statistics, RMSEA, RMSEA overlap, TLI and CFI across simulation conditions, they should not be utilized to capture the true model.

The second goal of the current study was to re-examine the capabilities of singlets, doublets and triplets χ^2/df ratios in identifying the correct response process when the non-convergence issues observed in Tay et al. (2011) study are resolved. As mentioned, results from the two studies by Tay et al. conflict, indicating that although test-level adjusted χ^2/df ratios doubles and triples could accurately capture the item response process when the underlying model was the GRM, this was not the case when the underlying model was the 2PL. Because the

2PL is a special case of the GRM for dichotomous data, findings regarding these two IRT models are expected to be consistent. In this study, when the number of iterations was increased to ensure the convergence of all replications, similar findings were observed across dichotomous and polytomous data. When the true response process was the ideal point model, all singlets, doublets and triplets χ^2/df ratios identified the correct model almost 100% of the time. In contrast, when the true model was the dominance model, the adjusted χ^2/df ratios could not differentiate the true model. Almost all χ^2/df ratios failed to point to the correct model, especially when the scale length was short. This is in stark contrast to the recommendations of Tay et al. (2011) to utilize χ^2/df ratio doubles and triples to identify the correct response process.

5.2 Study 2 Discussions

The first goal of Study 2 was to re-evaluate the Tay et al. study using the exact same study setting. Two questions were raised when reviewing the results from Tay et al.: a) a large number of non-convergences were observed when the estimation was based on the polytomous responses, but almost no non-converged replications were found when the items were dichotomously scored; and b) test-level doublets and triplets adjusted χ^2/df ratios could identify the true model accurately when the items were polytomously scored; however, these statistics performed much worse than when using the dichotomous data. Theoretically, the 2PL is simply a special case of the GRM for dichotomous data, meaning the GRM is a generalization of the 2PL to polytomous item response data (Embretson & Reise, 2000). Therefore, from a theoretical perspective, findings regarding these two IRT models should be highly consistent.

When the Tay et al. study was replicated in Study 2 using the same study design, the results in Study 2 were somewhat consistent with the Tay et al. study in that a large number of unsuccessful replications were found when the GGUM was fit to the simulated dominance-based

data regardless of the response types. In addition, the results in Study 2 indicated that the test-level singlets, doublets and triplets adjusted χ^2/df ratios could identify the true model almost 100% of the time when the true model was the ideal point model, whereas they performed much worse when the underlying true model was the dominance model, especially when sample size was relatively small. Unlike the inconsistent findings observed types in the Tay et al. study, the results in Study 2 were highly similar across the response types. Overall, the erratic findings in the Tay et al. study may be due to the aforementioned non-convergences.

The second goal of Study 2 was to examine the capabilities of identified model-data fit statistics when the model was not fully converged. Even though a large number of contaminated replications were observed, the overall impact on the capabilities of these tested model-data fit statistics did not vary much. As found in Study 1, no one universal model-data fit statistic can accurately identify the true response model 100% of the time; however, among all tested model-data fit statistics, the LL, AIC, and BIC were better model-data fit statistics for identifying the true underlying model regardless of the response types. These three model fit statistics performed especially well when the sample size was large enough with a reasonable scale length. The M_2 statistics, RMSEA, TLI, CFI, and adjusted χ^2/df ratios only performed well under certain response models, and therefore they should not be used to identify the true model. Also, the RMSEA overlap failed to differentiate the true model across all simulation conditions. These results are consistent with the findings in Study 1 even when the models were full converged.

5.3 General Discussions

Recent research has differentiated the dominance and the ideal point response processes. It remains vital for researchers and practitioners to correctly identify the response process in self-report data before proceeding to further statistical investigations. As suggested by current

research (e.g., Carter et al., 2014, 2017), failure to recognize the underlying item response process can impact both statistical outcomes and substantive conclusions.

Importantly, across all simulation conditions, the results of these two studies suggest that the LL, AIC and BIC outperformed other model-data fit statistics. When the sample size and scale length were sufficient, the LL, AIC, and BIC were able to differentiate the true model almost completely accurately. However, mixed results were found for the LL, AIC, and BIC when the data was not sufficient (e.g., N = 250 with a 15-item scale) when the true model was the GGUM. Because the GGUM is one of the most parameter-heavy IRT models, it is advisable to ensure the data is sufficient before calculating the LL, AIC and BIC values to identify the response process.

In addition, even though the M_2 statistics could differentiate the response process more accurately when the true model was the ideal point model, it performed extremely poorly when the true model was the dominance model. Similar patterns were found in the RMSEA, RMSEA overlap, TLI, and CFI. Surprisingly, these findings remained similar even when the results were contaminated with the non-converged estimations.

In the Tay et al. study, they concluded that comparing relative fits using test-level doublets and triplets adjusted χ^2/df ratios almost always pointed to the correct IRT model. However, findings from the current studies contradict these results. When the non-convergence issues were removed completely from the Tay et al. (2011) design by raising the number of iterations during the calibration, it became clear that the singlets, doublets, and triplets adjusted χ^2/df ratios could not assess the relative fit of ideal point and dominance models. The overall χ^2/df ratios failed to recognize the true model when the underlying model was the dominance model, whereas χ^2/df ratios identified the true model nearly 100% of the time when the

underlying model was the ideal point. Given the mixed findings across two response types, I do not recommend the use of the χ^2/df ratios to indicate the response process.

In summary, these simulations show that no universal model-data fit statistics can be used to differentiate the underlying response process of the data. However, the LL, AIC, and BIC can be used to assist with such decisions, but users should be aware that differences in AIC and BIC were generally small.

5.4 Limitations

Although these simulations have important implications in identifying the underlying response process of the data, these two studies still have limitations reducing the generalizability of the results. First, in order to validate the Tay et al. (2011) study, these two simulation studies followed the study design of Tay et al. closely. The only discrepancy was that Tay et al. used the FORESCORE software (Williams & Levin, 1993) to compute the doubles and triples adjusted χ^2/df ratios, while in the current study, an R program was developed to mimic the calculation of the FORSCORE. According to the documentation of the FORSCORE, when calculating the doublets and triplets, if the expected frequency was less than five, the number in the cell would be collapsed with a number in another cell and the degree of freedom would be reduced by one. Because the FORSCORE was poorly documented and the author was not able to explain the procedure (F. Drasgow, personal communication, July 8, 2016), this feature was not included in the current R program for calculating the adjusted χ^2/df ratios. The R code used in the current study can be found in Appendix A and can be investigated and utilized by future researchers. Future research should seek to further establish and verify a unified approach to calculating the adjusted χ^2/df ratios.

Secondly, even though the most popular model-data fit statistics were evaluated when identifying the response process, there are many other model-data fit statistics that can be used to investigate the true underlying model in the field of psychological measurement. As it is vital for researchers and practitioners to correctly identify the response process in self-report data before proceeding to further statistical investigations, future researchers are suggested to investigate and develop other model-fit statistics to identify the true response process. Some less popular model-data fit statistics may be appropriate to use to identify the true response model. Future research on these topics is highly encouraged.

CHAPTER 6

6 CONCLUSION

I believe these simulation results have important implications for psychological measurement research as well as practice. Even though the findings suggest that no one universal model-data fit statistic can be used to recognize the correct underlying model of the response data across all conditions, the LL, AIC, and BIC stood out as more suitable model-data fit statistics to identify the response process, especially when the sample size and scale length are sufficient. Moreover, findings suggest that the use of the M_2 statistics, RMSEA, RMSEA overlap, TLI, CFI, and adjusted χ^2/df ratios to identify the true response model is not appropriate. From a practical standpoint, given the importance of involving the psychological assessments across many stages in current organizational development, misidentifying the true underlying structure of the data may be detrimental to the people-related decisions. I encourage more research that enables a clearer understanding and identification of the response process. It is critical that we understand the underlying measurement model a psychological assessment to better access and interpret the results from such assessment.

References:

Bentler, P.M. (1990), Comparative Fit Indexes in Structural Models, Psychological Bulletin, 107 (2), 238-46.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88(3), 588-606.

Birnbaum, A. (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, 397-479.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62–83.

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Cai, L. (2017). flexMIRT R version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. British Journal of Mathematical and Statistical Psychology, 66, 245–276.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work Satisfaction Scale. Personality and Individual Differences, 49(7), 743–748. https://doi.org/10.1016/j.paid.2010.06.019

Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M.-C., & Delgado, K. M. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: how theoretically appropriate measurement makes an empirical difference. The Journal of Applied Psychology, 99(4), 564–586. https://doi.org/10.1037/a0034688

Carter, N. T., Dalal, D. K., Guan, L., LoPilato, A. C., & Withrow, S. A. (2017). Item response theory scoring and the detection of curvilinear relationships. Psychological Methods, 22(1), 191–203. https://doi.org/10.1037/met0000101

Carter, N. T., Guan, L., Maples, J., Williamson, R. L., & Miller, J. D. (2016). The downsides of high conscientiousness for psychological wellbeing: The role of obsessive compulsive tendencies. Journal of Personality.84(4), 510-522. https://doi.org/10.1111/jopy.12177

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chernyshenko, O. S., 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(1), 88–106. https://doi.org/10.1037/1040-3590.19.1.88

Cho, S., Drasgow, F., & Cao, M. (2015a). An investigation of emotional intelligence measures using item response theory. Psychological Assessment, 27(4), 1241–1252. https://doi.org/10.1037/pas0000132

Cho, S., Drasgow, F., & Cao, M. (2015b). An investigation of emotional intelligence measures using item response theory. Psychological Assessment, 27(4), 1241–1252. https://doi.org/10.1037/pas0000132

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

Cooper, L. G., & Nakanishi, M. (1983). Two logit models for external analysis of preferences. Psychometrika, 48(4), 607–620. https://doi.org/10.1007/BF02293883

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. Psychometrika, 42(4), 523–548. https://doi.org/10.1007/BF02295977

deLeeuw, J. (1992). Introduction to Akaike (1973) Information Theory and an Extension of the Maximum Likelihood Principle BT - Breakthroughs in Statistics: Foundations and Basic Theory. In S. Kotz & N. L. Johnson (Eds.) (pp. 599–609). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-0919-5_37

Drasgow, F., Levine, M. V, Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting Polyotomous Item Response Theory Models to Multiple-Choice Tests. Applied Psychological Measurement, 19(2), 143–165. https://doi.org/10.1177/014662169501900203

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

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Item Response Theory for Psychologists. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Grant, A. M., & Schwartz, B. (2011). Too Much of a Good Thing: The Challenge and Opportunity of the Inverted U. Perspectives on Psychological Science : A Journal of the Association for Psychological Science, 6(1), 61–76. https://doi.org/10.1177/1745691610393523

Guttman, L.A. (1950). The basis for scalogram analysis. In Stouffer, S.A., Guttman, L.A., & Schuman, E.A., Measurement and prediction. Volume 4 of Studies in social psychology in world war II. Princeton: Princeton University Press.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. Applied Psychological Measurement, 20(2), 101–125. https://doi.org/10.1177/014662169602000201

Hu, L.-T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA, US: Sage Publications, Inc.

Hu, L. T., & Bentler, P. M. (1999). Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. Structural Equation Modeling. Retrieved from http://proxyremote.galib.uga.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000208063500001&site=eds-live

Kamakura, W. A., & Srivastava, R. K. (1986). An Ideal-Point Probabilistic Choice Model for Heterogeneous Preferences. Marketing Science, 5(3), 199–218. https://doi.org/10.1287/mksc.5.3.199

Kenny, D. A., Kaniskan, B. and McCoach, D. B. The performance of RMSEA in models with small degrees of freedom. Sociological Methods & Research, 2014, 0049124114543236.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. Annals of the Institute of Statistical Mathematics, 18, 75–86.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. Ann. Math. Statist., 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: curvilinear relationships between personality traits and job performance. The Journal of Applied Psychology, 96(1), 113–133. https://doi.org/10.1037/a0021016

Ling, Y., Zhang, M., Locke, K. D., Li, G., & Li, Z. (2016). Examining the Process of Responding to Circumplex Scales of Interpersonal Values Items: Should Ideal Point Scoring Methods Be Considered? Journal of Personality Assessment, 98(3), 310–318. https://doi.org/10.1080/00223891.2015.1077852

Ling, Y., Zhang, M., Locke, K. D., Li, G., Li, Z., Ling, Y., … Li, Z. (2016). Examining the Process of Responding to Circumplex Scales of Interpersonal Values Items : Should Ideal Point Scoring Methods Be Considered ? Examining the Process of Responding to Circumplex Scales of Interpersonal Values Items : Should Ideal Point Scoring, 3891(April). https://doi.org/10.1080/00223891.2015.1077852

Liu, Y. & Tian, W. & Xin, T. (2016). An Application of M2 Statistic to Evaluate the Fit of Cognitive Diagnostic Models, Journal of Educational and Behavioral Statistics, , vol. 41(1), 3-26.

Lord, C. (1999). The ADOS-G (Autism Diagnostic Observation Schedule-Generic). Santa Monica, CA: Western Psychological Services.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Methods, 1(2), 130-149.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a Mixture Item Response Theory Model to Personality Questionnaire Data: Characterizing Latent Classes and Investigating Possibilities for Improving Prediction. Applied Psychological Measurement, 32(8), 611–631. https://doi.org/10.1177/0146621607312613

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. Journal of the American

Statistical Association, 100, 1009–1020

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. Personnel Psychology, 60(4), 995-1027. DOI: 10.1111/j.1744-6570.2007.00099.x

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Systematic Biology, 53(5), 793–808. https://doi.org/10.1080/10635150490522304

R Development Core Team (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Reise, S. P., & Waller, N. G. (1990). Fitting the Two-Parameter Model to Personality Data. Applied Psychological Measurement, 14(1), 45–58. https://doi.org/10.1177/014662169001400105

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. Psychometrika, 61, 509–528.

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. Sociological Methodology, 29, 81–111.

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(1), 3–32. https://doi.org/10.1177/01466216000241001

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(2), 192–207. https://doi.org/10.1177/01421602026002006

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2004). GGUM2004: A Windows based program to estimate parameters in the generalized graded unfolding model. Applied Psychological Measurement, 30, 64–65.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 34(4, Pt. 2), 100.

Schwarz, Gideon E. (1978). "Estimating the dimension of a model". Annals of Statistics 6 (2): 461–464. doi:10.1214/aos/1176344136.

Stark, S. (2001). MODFIT: A computer program for model-data fit. Unpublished manuscript. University of Illinois at Urbana–Champaign.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining Assumptions About Item Responding in Personality Assessment: Should Ideal Point Methods Be Considered for Scale Development and Scoring? Journal of Applied Psychology, 91(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Tay, L., Ali, U. S., 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(4), 280–295. https://doi.org/10.1177/0146621610390674

Tay, L., Ali, U. S. U. S., Drasgow, F., & Williams, B. Fitting irt models to dichotomous and polytomous data: Assessing the relative model–data fit of ideal point and dominance models., 35 Applied Psychological Measurement § (2011). Tay, Louis: Department of Psychology, University of Illinois at Urbana-Champaign, 603 E Daniel Street, Champaign, IL, US, 61820, Sage Publications. https://doi.org/10.1177/0146621610390674

Tay, L., & Drasgow, F. (2012). Theoretical, Statistical, and Substantive Issues in the Assessment

of Construct Dimensionality: Accounting for the Item Response Process. Organizational Research Methods, 15(3), 363–384. https://doi.org/10.1177/1094428112439709

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: are dominance models ideal? The Journal of Applied Psychology, 94(5), 1287–1304. https://doi.org/10.1037/a0015899

Tellegen, A. (1982). Brief manual for the Multidimensional Personality Questionnaire. Unpublished manuscript, University of Minnesota, Minneapolis

Thurstone, L. L. (1927). A law of comparative judgment. Psychological Review. US: Psychological Review Company. https://doi.org/10.1037/h0070288

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1–10.

van den Wollenberg, A. L. (1982). Two new test statistics for the rasch model. Psychometrika, 47(2), 123–140. https://doi.org/10.1007/BF02296270

Williams, B. A., & Levine, M. V. (1993). FORSCORE: A computer program for nonparametric item response theory. Unpublished manuscript, University of Illinois at Urbana-Champaign.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. Ann. Math. Statist., 9(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Zagorsek, H., Stough, S. J., & Jaklic, M. (2006). Analysis of the Reliability of the Leadership Practices Inventory in the Item Response Theory Framework. International Journal of Selection and Assessment, 14(2), 180–191. https://doi.org/10.1111/j.1468-2389.2006.00343.x

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis.

Psychometrika, 39(3), 327–350. https://doi.org/10.1007/BF02291707

APPENDIX A: Example R Code for Calculating GGUM-Based Adjusted χ^2/df Ratios

R Code	Calculating the singlets, doublets, and triplets adjusted χ^2/df ratios

```
All_Singlets <- matrix(ncol = 4, nrow = rep)
All_Doublets <- matrix(ncol = 4, nrow = rep)
All_Triplets <- matrix(ncol = 4, nrow = rep)

for (j in 1: rep){
  Response <- read.csv(paste("data_", (2*j-1), ".csv", sep = ""), header = T)
  Parameters <- read.csv(paste("item_parameters_ggum_", (2*j-1),".csv", sep =
""), header = T)
  Parameters <- cbind(Parameters[,2:3],(Parameters[,4:7]*(-1)))

  ########################Calclulate prob for each option ############

  for(i in 0:(Options-1)) {
    assign(paste("ALLOption", i, sep = ""), matrix(nrow = length(Density), ncol =
ITEMS))
  }

  for (i in 1:ITEMS) {

    A <- Parameters[i,1]
    D <- Parameters[i,2]
    T1 <- Parameters[i,3]
    T2 <- Parameters[i,4]
    T3 <- Parameters[i,5]
    T4 <- Parameters[i,6]

    Deno <- (exp(A*(0*(Quad-D)-0))+exp(A*((2*4+1-0)*(Quad-D)-
0)))+(exp(A*(1*(Quad-D)-T1))+exp(A*((2*4+1-1)*(Quad-D)-T1)))+
      (exp(A*(2*(Quad-D)-sum(T1,T2)))+exp(A*((2*4+1-2)*(Quad-D)-
sum(T1,T2))))+
      (exp(A*(3*(Quad-D)-sum(T1,T2,T3)))+exp(A*((2*4+1-3)*(Quad-D)-
sum(T1,T2,T3))))+
      (exp(A*(4*(Quad-D)-sum(T1,T2,T3,T4)))+exp(A*((2*4+1-4)*(Quad-D)-
sum(T1,T2,T3,T4))))

    Prob0 <- (exp(A*(0*(Quad-D)-0))+exp(A*((2*4+1-0)*(Quad-D)-0)))/Deno
    Prob1 <- (exp(A*(1*(Quad-D)-T1))+exp(A*((2*4+1-1)*(Quad-D)-T1)))/Deno
    Prob2 <- (exp(A*(2*(Quad-D)-sum(T1,T2)))+exp(A*((2*4+1-2)*(Quad-D)-
sum(T1,T2))))/Deno
    Prob3 <- (exp(A*(3*(Quad-D)-sum(T1,T2,T3)))+exp(A*((2*4+1-3)*(Quad-
D)-sum(T1,T2,T3))))/Deno
    Prob4 <- (exp(A*(4*(Quad-D)-sum(T1,T2,T3,T4)))+exp(A*((2*4+1-
4)*(Quad-D)-sum(T1,T2,T3,T4))))/Deno
```

```
    ALLOption0[,i] <- Prob0
    ALLOption1[,i] <- Prob1
    ALLOption2[,i] <- Prob2
    ALLOption3[,i] <- Prob3
    ALLOption4[,i] <- Prob4


  }


#####################Singlets#######################################

  Option <- function (ResponseOption) {

    Observed <- matrix(nrow = 1 ,ncol = ITEMS)
    Expected <- colSums(get(paste("ALLOption", ResponseOption, sep =
""))*Density)* 0.1 * N

    for (p in 1:ITEMS) {

      Observed[p] = c(length(which(Response[, p] == ResponseOption)))
    }

    Chi <- ((Observed-Expected)^2)/Expected
    return(Chi)
  }

  ChiSqrSinglets <- Option(0) + Option(1) + Option(2) + Option(3) + Option(4)
  ChiSqrDFSinglets <- ChiSqrSinglets/(Options-1)
  AdjustedSinglets <- (Options-1) + 3000*(ChiSqrSinglets-(Options-1))/N
  AdjustedSinglets <- ifelse(AdjustedSinglets < 0, 0, AdjustedSinglets)
  AdjustedDFSinglets <- AdjustedSinglets/(Options-1)

  Singlets <- t(rbind(ChiSqrSinglets,ChiSqrDFSinglets, AdjustedSinglets,
AdjustedDFSinglets))
  colnames(Singlets) <- c("ChiSqrSinglets","ChiSqrDFSinglets",
"AdjustedSinglets", "AdjustedDFSinglets")
  Final_Singlets <- t(colMeans(Singlets, na.rm = TRUE))


#####################Doublets#######################################

  # create a two-way contigency table
  Doublets <- function(item1, item2) {
    Observed <- matrix(nrow = Options, ncol = Options)
```

```
    Expected <- matrix(nrow = Options, ncol = Options)

    for (m in 0:(Options-1)) {
      for (n in 0:(Options-1)) {

        Expected[(m+1),(n+1)] <- sum((get(paste("ALLOption", m, sep = ""))[,
item1]
                                      +get(paste("ALLOption", n, sep = ""))[,
item2])*Density)*0.1*N
        Observed[(m+1),(n+1)] <- c(length(which(Response[, item1] == m))) +
c(length(which(Response[, item2] == n)))
      }
    }

    ChiSqrDoublets <- sum(((Observed-Expected)^2)/Expected)
    return(ChiSqrDoublets)

  }

  ChiSqrDoublets <- data.frame()

  for (p in 1:(ITEMS-1)) {
    for (q in (p+1):ITEMS) {

      Doublets_all <- Doublets(p, q)
      ChiSqrDoublets <- rbind(ChiSqrDoublets, Doublets_all)
    }
  }

  df <- Options*Options - 1
  ChiSqrDFDoublets <- t(ChiSqrDoublets/df)
  AdjustedDoublets <- df + 3000*(t(ChiSqrDoublets)-df)/N
  AdjustedDoublets <- ifelse(AdjustedDoublets < 0, 0, AdjustedDoublets)
  AdjustedDFDoublets <- AdjustedDoublets/df

  Doublets <- t(rbind(t(ChiSqrDoublets),ChiSqrDFDoublets, AdjustedDoublets,
AdjustedDFDoublets))
  colnames(Doublets) <- c("ChiSqrDoublets","ChiSqrDFDoublets",
"AdjustedDoublets", "AdjustedDFDoublets")
  Final_Doublets <- t(colMeans(Doublets, na.rm = TRUE))

  ######################Triplets#############################

  Triplets <- function(item1, item2, item3) {

    Expected <- data.frame()
```

```
  Observed <- data.frame()

  for (m in 0:(Options-1)) {
    for (n in 0:(Options-1)) {
      for (t in 0:(Options-1)) {

        Expected_calculate <- sum((get(paste("ALLOption", m, sep = ""))[,
item1]+
                                   get(paste("ALLOption", n, sep = ""))[, item2]+
                                   get(paste("ALLOption", t, sep = ""))[,
item3])*Density)*0.1*N
        Observed_calculate <- c(length(which(Response[, item1] == m)))+
          c(length(which(Response[, item2] == n)))+
          c(length(which(Response[, item3] == t)))

        Expected <- rbind(Expected, Expected_calculate)
        Observed <- rbind(Observed, Observed_calculate)

      }
    }
  }

  ChiSqrTriplets <- sum(((Observed-Expected)^2)/Expected)
  return(ChiSqrTriplets)

}

ChiSqrTriplets <- data.frame()

for (p in 1:(ITEMS-2)) {
  for (q in (p+1):(ITEMS-1)) {
    for (t in (q+1):ITEMS) {

      Triplets_all <- Triplets(p, q, t)
      ChiSqrTriplets <- rbind(ChiSqrTriplets, Triplets_all)
    }
  }
}

df <- Options*Options*Options-1
ChiSqrDFTriplets <- t(ChiSqrTriplets/df)
AdjustedTriplets <- df + 3000*(t(ChiSqrTriplets)-df)/N
AdjustedTriplets <- ifelse(AdjustedTriplets < 0, 0, AdjustedTriplets)
AdjustedDFTriplets <- AdjustedTriplets/df

Triplets <- t(rbind(t(ChiSqrTriplets),ChiSqrDFTriplets, AdjustedTriplets,
```

```
AdjustedDFTriplets))
  colnames(Triplets) <- c("ChiSqrTriplets","ChiSqrDFTriplets",
"AdjustedTriplets", "AdjustedDFTriplets")
  Final_Triplets <- t(colMeans(Triplets, na.rm = TRUE))


  All_Singlets[j, ] <- Final_Singlets
  All_Doublets[j, ] <- Final_Doublets
  All_Triplets[j, ] <- Final_Triplets
}

Singlets_cali <- t(colMeans(All_Singlets, na.rm = TRUE))
colnames(Singlets_cali) <- c("ChiSqrSinglets","ChiSqrDFSinglets",
"AdjustedSinglets", "AdjustedDFSinglets")
Doublets_cali <- t(colMeans(All_Doublets, na.rm = TRUE))
colnames(Doublets_cali) <- c("ChiSqrDoublets","ChiSqrDFDoublets",
"AdjustedDoublets", "AdjustedDFDoublets")
Triplets_cali <- t(colMeans(All_Triplets, na.rm = TRUE))
colnames(Triplets_cali) <- c("ChiSqrTriplets","ChiSqrDFTriplets",
"AdjustedTriplets", "AdjustedDFTriplets")
```

Table 1: Non-convergence rates by condition when fitting the GGUM to data arising from ideal point and dominance response processes

	Dichotomous data							
Data type	GGUM				2PL			
	Simulated data for 15-item scale across 200 replications							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	2%	0%	0%	0%	5%	1%	0%	0%
	Simulated data for 30-item scale across 200 replications							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	5.5%	0%	0%	0%	6.5%	1.5%	0%	0.5%
	Polytomous data							
Data type	GGUM				GRM			
	Simulated data for 5-item scale across 200 replications							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	**29%**	**24%**	**24.5%**	**22.5%**
	Simulated data for 10-item scale across 200 replications							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	**16%**	**10%**	**12%**	**6.5%**

Note: NC-R = non-convergence rate (i.e., the number of non-converged replications in each condition / the number of total replications in each condition).

Table 2: Dichotomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the two-parameter logistic model (2PL)

Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
2PL	15	250	C	-2099.28	4258.57	4364.21	90.13	90	0.01	0.00	0.03	1.03	0.96	2.66	1.75	1.49
			V	-2100.83	4261.65	4367.29	91.70	90	0.01	0.00	0.03	1.02	0.95	3.24	1.99	1.66
		500	C	-4274.53	8609.05	8735.49	91.15	90	0.01	0.00	0.02	1.00	0.98	2.64	1.77	1.52
			V	-4261.75	8583.51	8709.94	90.67	90	0.01	0.00	0.02	1.00	0.98	2.83	1.84	1.58
		1000	C	-8513.37	17086.74	17233.97	91.58	90	0.01	0.00	0.02	1.00	0.99	3.33	2.07	1.74
			V	-8515.17	17090.34	17237.57	90.85	90	0.01	0.00	0.02	1.00	0.99	3.48	2.13	1.78
		2000	C	-16954.80	33969.61	34137.64	89.94	90	0.00	0.00	0.01	1.00	1.00	2.81	1.90	1.63
			V	-16946.79	33953.59	34121.61	90.10	90	0.00	0.00	0.01	1.00	0.99	2.87	1.93	1.65
	30	250	C	-4177.90	8475.80	8687.08	406.09	405	0.01	0.00	0.02	1.00	0.98	2.63	1.74	1.48
			V	-4164.35	8448.70	8659.99	415.19	405	0.01	0.00	0.02	0.99	0.98	3.15	1.96	1.64
		500	C	-8413.85	16947.70	17200.57	412.16	405	0.01	0.00	0.02	1.00	0.99	2.70	1.81	1.55
			V	-8418.08	16956.15	17209.03	409.63	405	0.01	0.00	0.02	1.00	0.99	3.00	1.93	1.64
		1000	C	-16862.50	33845.00	34139.46	402.66	405	0.00	0.00	0.01	1.00	1.00	2.71	1.82	1.57
			V	-16871.19	33862.37	34156.84	403.83	405	0.00	0.00	0.01	1.00	1.00	2.87	1.89	1.61
		2000	C	-33621.96	67363.92	67699.97	404.22	405	0.00	0.00	0.01	1.00	1.00	2.72	1.85	1.59
			V	-33619.61	67359.22	67695.28	402.40	405	0.00	0.00	0.01	1.00	1.00	3.05	2.01	1.71
Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
GGUM	15	250	C	-2092.83	4275.66	4434.12	68.74	75	0.01	0.00	0.03	1.01	1.00	0.47	0.28	0.20
			V	-2094.05	4278.10	4436.57	69.09	75	0.01	0.00	0.03	1.01	1.00	6.27	2.75	1.80
		500	C	-4272.51	8635.01	8824.67	68.59	75	0.00	0.00	0.02	1.00	1.00	0.42	0.27	0.21
			V	-4258.91	8607.82	8797.48	68.34	75	0.00	0.00	0.02	1.00	1.00	3.32	1.60	1.14
		1000	C	-8522.33	17134.65	17355.50	70.64	75	0.00	0.00	0.01	1.00	1.00	0.67	0.46	0.38
			V	-8520.80	17131.59	17352.44	68.71	75	0.00	0.00	0.01	1.00	1.00	1.38	1.11	1.04
		2000	C	-16970.67	34031.35	34283.39	69.84	75	0.00	0.00	0.01	1.00	1.00	0.14	0.89	0.89
			V	-16964.58	34019.17	34271.21	69.13	75	0.00	0.00	0.01	1.00	1.00	1.16	1.03	1.00
	30	250	C	-4174.57	8529.14	8846.07	363.96	375	0.00	0.00	0.02	1.00	1.00	0.38	0.26	0.22
			V	-4165.33	8510.66	8827.59	372.52	375	0.01	0.00	0.02	1.00	1.00	2.40	1.46	1.21
		500	C	-8425.60	17031.21	17410.52	368.05	375	0.00	0.00	0.01	1.00	1.00	3.04	2.02	1.66
			V	-8429.68	17039.35	17418.67	366.06	375	0.00	0.00	0.01	1.00	1.00	12.45	6.12	4.34
		1000	C	-16926.20	34032.39	34474.09	368.48	375	0.00	0.00	0.01	1.00	1.00	1.24	1.09	1.04
			V	-16929.60	34039.20	34480.90	369.41	375	0.00	0.00	0.01	1.00	1.00	1.75	1.39	1.29
		2000	C	-33745.11	67670.22	68174.30	388.45	375	0.00	0.00	0.01	1.00	1.00	5.46	3.73	3.14
			V	-33696.96	67573.93	68078.01	378.98	375	0.00	0.00	0.01	1.00	1.00	20.09	12.16	9.98

Note: C = calibration sample; V = validation sample.

Table 3: Dichotomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the generalized graded unfolding model (GGUM)

Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
2PL	15	250	C	-2367.84	4795.69	4901.33	146.92	90	0.05	0.03	0.06	0.83	0.85	8.17	3.28	2.13
			V	-2368.92	4797.84	4903.48	139.47	90	0.04	0.03	0.06	0.84	0.86	15.00	7.80	6.04
		500	C	-4743.22	9546.43	9672.87	185.82	90	0.04	0.03	0.05	0.84	0.86	9.55	4.78	3.67
			V	-4742.29	9544.59	9671.02	186.59	90	0.04	0.03	0.05	0.84	0.86	12.45	6.81	5.47
		1000	C	-9548.82	19157.64	19304.87	328.94	90	0.05	0.04	0.06	0.80	0.83	9.90	5.64	4.60
			V	-9543.95	19147.90	19295.13	327.57	90	0.05	0.04	0.06	0.80	0.83	9.67	5.70	4.72
		2000	C	-19041.16	38142.31	38310.34	514.36	90	0.05	0.04	0.05	0.83	0.85	10.24	5.93	4.92
			V	-19048.84	38157.67	38325.70	520.87	90	0.05	0.04	0.05	0.82	0.85	9.23	5.45	4.57
	30	250	C	-4699.66	9519.32	9730.61	705.86	405	0.05	0.05	0.06	0.78	0.79	10.00	4.20	2.84
			V	-4696.81	9513.62	9724.91	703.05	405	0.05	0.05	0.06	0.79	0.80	17.09	8.92	6.92
		500	C	-9402.07	18924.14	19177.01	1049.48	405	0.06	0.05	0.06	0.78	0.79	10.32	5.36	4.14
			V	-9408.02	18936.03	19188.91	1055.41	405	0.06	0.05	0.06	0.77	0.79	13.02	7.24	5.82
		1000	C	-18797.88	37715.77	38010.23	1576.03	405	0.05	0.05	0.06	0.80	0.81	10.87	6.15	5.00
			V	-18794.07	37708.14	38002.61	1596.85	405	0.05	0.05	0.06	0.80	0.81	11.07	6.46	5.32
		2000	C	-37769.22	75658.44	75994.49	2912.20	405	0.05	0.05	0.06	0.77	0.79	11.57	6.79	5.62
			V	-37773.52	75667.03	76003.09	2946.55	405	0.06	0.05	0.06	0.77	0.79	10.16	6.07	5.07
Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
GGUM	15	250	C	-2345.73	4781.46	4939.92	81.01	75	0.02	0.00	0.04	0.99	0.98	0.32	0.22	0.19
			V	-2346.33	4782.65	4941.12	79.11	75	0.01	0.00	0.04	0.99	0.98	5.30	2.79	2.05
		500	C	-4705.89	9501.78	9691.44	79.02	75	0.01	0.00	0.03	1.00	0.99	0.88	0.76	0.72
			V	-4707.13	9504.25	9693.91	79.27	75	0.01	0.00	0.03	1.00	0.99	3.91	2.37	1.93
		1000	C	-9453.56	18997.11	19217.96	82.48	75	0.01	0.00	0.02	1.00	0.99	0.97	0.90	0.88
			V	-9459.13	19008.27	19229.12	82.41	75	0.01	0.00	0.02	1.00	0.99	2.39	1.66	1.45
		2000	C	-18919.00	37928.01	38180.05	103.52	75	0.01	0.00	0.02	0.99	0.99	1.17	1.08	1.05
			V	-18906.86	37903.71	38155.75	89.12	75	0.01	0.00	0.01	1.00	1.00	2.26	1.71	1.55
	30	250	C	-4609.15	9398.31	9715.24	389.80	375	0.01	0.00	0.02	0.99	0.99	0.61	0.41	0.35
			V	-4608.88	9397.76	9714.70	389.44	375	0.01	0.00	0.03	0.99	0.99	6.08	3.30	2.49
		500	C	-9227.59	18635.18	19014.49	400.65	375	0.01	0.00	0.02	0.99	0.99	1.14	0.93	0.87
			V	-9240.32	18660.64	19039.96	406.55	375	0.01	0.00	0.02	0.99	0.99	4.36	2.68	2.20
		1000	C	-18502.90	37185.79	37627.49	440.07	375	0.01	0.01	0.02	0.99	0.99	1.60	1.33	1.24
			V	-18476.49	37132.99	37574.69	438.61	375	0.01	0.01	0.02	0.99	0.99	4.13	2.83	2.45
		2000	C	-37149.79	74479.58	74983.66	519.88	375	0.01	0.01	0.01	0.99	0.99	1.60	1.36	1.29
			V	-37158.29	74496.57	75000.65	544.24	375	0.01	0.01	0.01	0.99	0.99	3.42	2.50	2.22

Note: C = calibration sample; V = validation sample.

Table 4: Dichotomous data: percentages of fit statistics across all replications pointing to the correct model

True Model	Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
												Singlets	Doublets	Triplets
2PL	15	250	C	6	92	100	0	57	45	49	53	5	4	2
			V	11	92	100	0	60	44	51	55	75	63	46
		500	C	28	97	100	0	51	39	45	47	4	4	3
			V	30	97	100	1	52	36	44	48	54	33	26
		1000	C	59	100	100	1	55	41	50	50	1	1	1
			V	62	100	100	2	58	46	43	48	1	1	1
		2000	C	68	100	100	0	56	40	45	48	0	0	0
			V	71	100	100	1	62	53	50	52	0	0	0
	30	250	C	38	100	100	1	51	47	40	41	2	2	2
			V	33	100	100	0	53	41	44	48	14	13	11
		500	C	55	100	100	0	55	46	45	49	23	23	23
			V	60	100	100	0	52	43	40	45	77	69	54
		1000	C	83	100	100	4	65	53	49	51	11	15	15
			V	77	100	100	5	64	49	50	54	11	14	15
		2000	C	70	99	99	6	76	60	58	59	23	23	23
			V	78	100	100	8	69	57	52	54	66	52	47
GGUM	15	250	C	81	66	24	98	93	7	96	96	100	98	95
			V	82	67	26	98	92	10	95	95	96	95	95
		500	C	79	72	46	100	99	1	100	100	99	99	99
			V	83	73	49	99	98	2	99	99	96	94	93
		1000	C	82	78	68	98	96	4	98	99	100	100	100
			V	84	81	70	99	99	0	99	100	98	97	97
		2000	C	78	75	72	99	98	2	99	99	100	100	100
			V	74	73	68	97	97	1	97	97	95	92	92
	30	250	C	94	81	53	100	100	0	100	100	100	99	99
			V	96	86	60	100	100	0	100	100	99	99	97
		500	C	95	93	76	100	100	0	100	100	100	100	100
			V	96	88	77	100	100	0	100	100	97	96	96
		1000	C	88	84	78	100	100	0	100	100	100	100	100
			V	88	85	81	100	100	0	100	100	94	92	92
		2000	C	91	91	89	100	99	1	99	99	100	100	100
			V	92	92	88	100	100	0	100	100	95	92	91

Note: C = calibration sample; V = validation sample.

Table 5: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true and the calibrated model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3827.33	7704.66	7810.02	-	-	-	-	-	-	-	2.41	1.98	1.87
		V	-3829.01	7708.01	7813.38	-	-	-	-	-	-	-	5.15	3.91	3.59
	1000	C	-7669.40	15388.80	15511.50	-	-	-	-	-	-	-	2.51	2.02	1.90
		V	-7665.94	15381.87	15504.56	-	-	-	-	-	-	-	5.48	4.03	3.70
	1500	C	-11476.73	23003.47	23136.30	-	-	-	-	-	-	-	2.37	1.97	1.87
		V	-11471.84	22993.69	23126.52	-	-	-	-	-	-	-	5.06	3.90	3.60
	2000	C	-15306.38	30662.77	30802.79	-	-	-	-	-	-	-	2.51	2.00	1.89
		V	-15308.17	30666.35	30806.37	-	-	-	-	-	-	-	5.50	3.99	3.66
10	500	C	-7589.91	15279.83	15490.56	-	-	-	-	-	-	-	2.71	2.04	1.92
		V	-7593.26	15286.50	15497.19	-	-	-	-	-	-	-	6.02	4.07	3.72
	1000	C	-15223.15	30546.31	30791.69	-	-	-	-	-	-	-	2.47	2.01	1.90
		V	-15220.60	30541.19	30786.58	-	-	-	-	-	-	-	5.39	4.05	3.71
	1500	C	-22859.37	45818.75	46084.41	-	-	-	-	-	-	-	2.31	1.94	1.84
		V	-22858.38	45816.77	46082.43	-	-	-	-	-	-	-	4.91	3.81	3.51
	2000	C	-30468.16	61036.33	61316.37	-	-	-	-	-	-	-	2.34	1.97	1.87
		V	-30478.53	61057.07	61337.11	-	-	-	-	-	-	-	5.03	3.91	3.60
15	500	C	-11360.12	22870.24	23186.34	44.58	45	0.01	0.00	0.03	1.05	0.92	2.39	1.97	1.86
		V	-11361.87	22873.72	23189.78	45.73	45	0.01	0.00	0.03	1.02	0.89	5.16	3.90	3.58
	1000	C	-22728.37	45606.73	45974.76	46.35	45	0.01	0.00	0.02	0.99	0.95	2.39	1.98	1.87
		V	-22743.81	45637.63	46005.71	44.66	45	0.01	0.00	0.02	1.01	0.96	5.13	3.92	3.61
	1500	C	-34192.97	68535.94	68934.43	45.70	45	0.01	0.00	0.02	1.00	0.97	2.41	1.99	1.88
		V	-34187.57	68525.14	68923.63	46.65	45	0.01	0.00	0.02	0.99	0.97	5.23	3.97	3.65
	2000	C	-45555.33	91260.66	91680.72	44.97	45	0.00	0.00	0.01	1.00	0.98	2.36	1.95	2.14
		V	-45562.49	91274.98	91695.05	45.16	45	0.00	0.00	0.01	1.00	0.98	5.07	3.83	3.53
20	500	C	-15125.48	30450.96	30872.42	111.08	110	0.01	0.00	0.02	1.00	0.95	2.48	1.98	1.87
		V	-15125.88	30451.74	30873.16	109.43	110	0.01	0.00	0.02	1.02	0.96	5.38	3.92	3.59
	1000	C	-30254.02	60708.04	61198.82	109.00	110	0.00	0.00	0.01	1.00	0.98	2.45	2.00	1.88
		V	-30254.11	60708.21	61198.99	109.23	110	0.00	0.00	0.01	1.00	0.98	5.28	3.95	3.62
	1500	C	-45382.31	90964.61	91495.93	111.13	110	0.00	0.00	0.01	1.00	0.99	2.36	1.96	1.86
		V	-45388.67	90977.33	91508.65	112.24	110	0.00	0.00	0.01	1.00	0.99	5.05	3.88	3.57
	2000	C	-60595.13	121390.25	121950.34	110.00	110	0.00	0.00	0.01	1.00	0.99	2.36	1.97	1.87
		V	-60576.55	121353.10	121913.19	110.81	110	0.00	0.00	0.01	1.00	0.99	5.08	3.90	3.60

Note: C = calibration sample; V = validation sample.

Table 6: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true model was the graded response model (GRM) and the calibrated model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3827.84	7715.69	7842.13	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-3831.04	7722.08	7848.52	-	-	-	-	-	-	-	8.17	5.58	5.07
	1000	C	-7673.14	15406.28	15553.51	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-7669.16	15398.33	15545.56	-	-	-	-	-	-	-	3.97	2.87	2.66
	1500	C	-11483.38	23026.76	23186.15	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-11479.54	23019.08	23178.47	-	-	-	-	-	-	-	3.12	2.32	2.18
	2000	C	-15318.23	30696.47	30864.50	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-15319.71	30699.42	30867.44	-	-	-	-	-	-	-	2.48	1.94	1.85
10	500	C	-7605.55	15331.09	15583.97	-	-	-	-	-	-	-	1.98	1.55	1.36
		V	-7608.81	15337.59	15590.43	-	-	-	-	-	-	-	9.48	6.71	6.17
	1000	C	-15257.17	30634.34	30928.81	-	-	-	-	-	-	-	1.65	1.36	1.26
		V	-15255.60	30631.19	30925.66	-	-	-	-	-	-	-	5.72	4.28	4.03
	1500	C	-22916.11	45952.22	46271.01	-	-	-	-	-	-	-	1.48	1.25	1.17
		V	-22918.07	45956.14	46274.93	-	-	-	-	-	-	-	4.45	3.51	3.34
	2000	C	-30550.61	61221.23	61557.28	-	-	-	-	-	-	-	1.71	1.45	1.39
		V	-30566.67	61253.34	61589.39	-	-	-	-	-	-	-	4.23	3.43	3.29
15	500	C	-11411.03	23002.05	23381.36	27.52	30	0.01	0.00	0.03	0.21	0.81	4.04	3.17	2.79
		V	-11411.52	23003.02	23382.30	28.27	30	0.01	0.00	0.03	1.08	0.85	11.54	8.48	7.87
	1000	C	-22825.45	45830.88	46272.52	30.51	30	0.01	0.00	0.02	0.74	0.91	3.75	3.11	2.89
		V	-22840.06	45860.12	46301.82	29.93	30	0.01	0.00	0.02	1.08	0.93	7.51	5.87	5.56
	1500	C	-34362.65	68905.30	69383.49	32.43	30	0.01	0.00	0.02	0.94	0.90	4.29	3.62	3.43
		V	-34336.07	68852.14	69330.33	29.65	30	0.01	0.00	0.02	1.01	0.95	7.36	6.01	5.75
	2000	C	-45764.99	91709.98	92214.06	31.35	30	0.01	0.00	0.02	0.91	0.94	3.44	2.93	2.81
		V	-45757.00	91693.99	92198.07	32.51	30	0.01	0.00	0.02	0.95	0.95	5.93	4.90	4.70
20	500	C	-15196.39	30632.78	31138.53	90.36	90	0.01	0.00	0.02	0.97	0.83	4.13	3.21	2.81
		V	-15207.98	30655.95	31161.66	89.45	90	0.01	0.00	0.02	0.75	0.82	11.52	8.54	7.95
	1000	C	-30396.39	61032.78	61621.71	90.15	90	0.01	0.00	0.02	1.06	0.93	3.75	3.10	2.88
		V	-30397.58	61035.15	61624.08	94.27	90	0.01	0.00	0.02	0.95	0.88	7.61	5.94	5.62
	1500	C	-45631.64	91503.28	92140.87	97.50	90	0.01	0.00	0.01	0.92	0.90	4.39	3.71	3.53
		V	-45647.74	91535.48	92173.07	96.47	90	0.01	0.00	0.01	0.94	0.90	7.48	6.09	5.82
	2000	C	-61003.23	122246.47	122918.58	102.20	90	0.01	0.00	0.01	0.89	0.89	5.78	4.93	4.73
		V	-60959.66	122159.31	122831.42	100.85	90	0.01	0.00	0.01	0.89	0.90	8.48	7.09	6.81

Note: C = calibration sample; V = validation sample.

Table 7: Polytomous data: percentages of fit statistics across all replications pointing to the correct model when the true model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
											Singlets	Doublets	Triplets
5	500	C	58	90	100	-	-	-	-	-	0	0	0
		V	62	96	100	-	-	-	-	-	79	65	62
	1000	C	70	99	100	-	-	-	-	-	0	0	0
		V	77	99	100	-	-	-	-	-	27	26	28
	1500	C	87	100	100	-	-	-	-	-	0	0	0
		V	83	99	100	-	-	-	-	-	13	12	11
	2000	C	98	100	100	-	-	-	-	-	0	0	0
		V	94	100	100	-	-	-	-	-	4	3	4
10	500	C	77	100	100	-	-	-	-	-	22	19	17
		V	74	99	100	-	-	-	-	-	70	77	76
	1000	C	84	100	100	-	-	-	-	-	17	17	17
		V	89	100	100	-	-	-	-	-	42	42	43
	1500	C	96	100	100	-	-	-	-	-	15	15	14
		V	95	100	100	-	-	-	-	-	25	27	29
	2000	C	100	100	100	-	-	-	-	-	17	17	17
		V	98	100	100	-	-	-	-	-	17	17	18
15	500	C	88	100	100	5	64	55	47	69	41	39	38
		V	89	100	100	5	68	63	40	31	96	89	89
	1000	C	92	100	100	5	69	60	46	69	47	46	45
		V	91	100	100	5	65	59	42	27	59	56	58
	1500	C	99	100	100	8	64	50	37	64	48	48	48
		V	100	100	100	14	75	63	54	44	51	53	55
	2000	C	98	100	100	9	72	58	47	73	47	47	47
		V	99	100	100	10	68	58	45	35	44	47	49
20	500	C	95	100	100	8	67	59	45	77	49	46	44
		V	86	100	100	8	65	60	50	44	95	87	86
	1000	C	97	100	100	12	82	71	52	89	59	60	57
		V	97	100	100	8	71	66	44	41	68	69	69
	1500	C	99	100	100	17	75	61	55	81	53	54	54
		V	99	100	100	18	68	59	54	50	61	63	65
	2000	C	99	100	100	23	75	63	61	80	70	70	70
		V	99	100	100	27	78	69	59	58	67	71	72

Note: C = calibration sample; V = validation sample.

Table 8: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true and the calibrated model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3356.93	6773.82	6900.18	-	-	-	-	-	-	-	0.57	0.57	0.56
		V	-3350.34	6760.68	6887.12	-	-	-	-	-	-	-	1.61	1.40	1.36
	1000	C	-6766.70	13593.41	13740.64	-	-	-	-	-	-	-	0.75	0.75	0.75
		V	-6760.16	13580.32	13727.56	-	-	-	-	-	-	-	1.24	1.15	1.13
	1500	C	-10176.73	20413.46	20572.85	-	-	-	-	-	-	-	1.04	1.02	1.02
		V	-10171.51	20403.02	20562.42	-	-	-	-	-	-	-	1.37	1.29	1.27
	2000	C	-13549.58	27159.15	27327.18	-	-	-	-	-	-	-	0.88	0.88	0.88
		V	-13565.14	27190.27	27358.30	-	-	-	-	-	-	-	1.15	1.10	1.09
10	500	C	-6694.08	13508.16	13761.04	-	-	-	-	-	-	-	1.31	1.20	1.18
		V	-6686.95	13493.91	13746.78	-	-	-	-	-	-	-	2.25	1.96	1.90
	1000	C	-13262.99	26645.97	26940.44	-	-	-	-	-	-	-	1.71	1.59	1.57
		V	-13233.81	26587.62	26882.08	-	-	-	-	-	-	-	2.54	2.27	2.22
	1500	C	-20078.64	40277.27	40596.07	-	-	-	-	-	-	-	1.85	1.72	1.69
		V	-20096.75	40313.50	40632.30	-	-	-	-	-	-	-	2.17	1.98	1.94
	2000	C	-26416.86	52953.71	53289.76	-	-	-	-	-	-	-	1.48	1.41	1.39
		V	-26389.52	52899.04	53235.10	-	-	-	-	-	-	-	1.72	1.60	1.57
15	500	C	-9733.04	19646.06	20025.33	204.65	30	0.08	0.06	0.09	0.93	0.95	1.15	1.08	1.06
		V	-9763.56	19707.11	20086.43	213.73	30	0.08	0.07	0.10	0.92	0.95	2.46	2.06	2.00
	1000	C	-19490.87	39161.74	39603.44	319.80	30	0.07	0.06	0.09	0.93	0.95	1.31	1.24	1.23
		V	-19498.43	39176.85	39618.55	309.75	30	0.07	0.06	0.08	0.93	0.95	1.75	1.60	1.56
	1500	C	-29254.97	58689.94	59168.13	501.63	30	0.07	0.06	0.08	0.93	0.96	1.46	1.39	1.37
		V	-29277.43	58734.87	59213.06	556.48	30	0.07	0.07	0.08	0.92	0.95	1.75	1.62	1.59
	2000	C	-38998.18	78176.36	78680.44	613.12	30	0.07	0.06	0.08	0.93	0.95	1.42	1.35	1.34
		V	-39010.34	78200.69	78704.77	699.50	30	0.07	0.07	0.08	0.92	0.94	1.64	1.53	1.51
20	500	C	-13032.70	26305.40	26811.16	591.99	90	0.09	0.08	0.10	0.93	0.94	1.65	1.51	1.47
		V	-12994.93	26229.86	26735.61	565.28	90	0.08	0.08	0.09	0.93	0.94	2.47	2.17	2.10
	1000	C	-25971.46	52182.92	52771.85	1099.49	90	0.09	0.08	0.09	0.94	0.94	1.84	1.71	1.68
		V	-25902.29	52044.58	52633.52	1020.64	90	0.08	0.08	0.09	0.94	0.94	2.67	2.36	2.30
	1500	C	-38881.18	78002.36	78639.94	1290.62	90	0.08	0.07	0.08	0.94	0.95	1.76	1.64	1.62
		V	-38852.37	77944.73	78582.32	1337.63	90	0.08	0.07	0.08	0.94	0.95	2.04	1.88	1.84
	2000	C	-51663.05	103566.10	104238.21	1621.78	90	0.07	0.07	0.08	0.94	0.95	1.74	1.64	1.61
		V	-51646.76	103533.51	104205.62	1611.80	90	0.07	0.07	0.08	0.94	0.95	1.96	1.81	1.78

Note: C = calibration sample; V = validation sample.

Table 9: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true model was the generalized graded unfolding model (GGUM) and the calibrated model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3424.28	6898.53	7003.81	-	-	-	-	-	-	-	6.60	4.21	3.76
		V	-3418.39	6886.79	6992.15	-	-	-	-	-	-	-	8.03	5.19	4.69
	1000	C	-6915.65	13881.29	14003.99	-	-	-	-	-	-	-	6.11	4.16	3.80
		V	-6914.00	13878.01	14000.70	-	-	-	-	-	-	-	7.08	4.75	4.34
	1500	C	-10403.69	20857.39	20990.22	-	-	-	-	-	-	-	7.61	4.90	4.39
		V	-10409.97	20869.93	21002.76	-	-	-	-	-	-	-	8.24	5.22	4.67
	2000	C	-13818.75	27687.51	27827.53	-	-	-	-	-	-	-	6.49	4.35	3.96
		V	-13822.49	27694.99	27835.01	-	-	-	-	-	-	-	7.11	4.71	4.27
10	500	C	-6830.10	13760.21	13970.94	-	-	-	-	-	-	-	7.81	4.97	4.45
		V	-6832.99	13765.98	13976.71	-	-	-	-	-	-	-	9.33	6.01	5.40
	1000	C	-13486.35	27072.71	27318.09	-	-	-	-	-	-	-	10.23	5.90	5.04
		V	-13482.91	27065.82	27311.21	-	-	-	-	-	-	-	8.99	5.65	5.03
	1500	C	-20327.26	40754.51	41020.17	-	-	-	-	-	-	-	7.93	5.16	4.63
		V	-20314.83	40729.66	40995.33	-	-	-	-	-	-	-	8.42	5.45	4.90
	2000	C	-27036.58	54173.16	54453.21	-	-	-	-	-	-	-	8.24	5.20	4.63
		V	-27030.52	54161.05	54441.09	-	-	-	-	-	-	-	8.63	5.46	4.86
15	500	C	-10056.65	20263.27	20579.33	481.45	45	0.13	0.12	0.14	0.87	0.91	8.91	5.65	5.02
		V	-10054.81	20259.62	20575.71	448.46	45	0.12	0.11	0.14	0.88	0.91	8.63	5.46	4.86
	1000	C	-20121.28	40392.55	40760.63	862.31	45	0.13	0.12	0.13	0.88	0.91	8.66	5.54	4.92
		V	-20123.10	40396.20	40764.28	920.23	45	0.13	0.12	0.14	0.88	0.91	9.71	6.16	5.47
	1500	C	-30223.82	60597.65	60996.14	1274.96	45	0.13	0.12	0.13	0.88	0.91	9.23	5.77	5.08
		V	-30225.87	60601.74	61000.23	1284.86	45	0.13	0.12	0.13	0.88	0.91	9.82	6.12	5.39
	2000	C	-40242.75	80635.51	81055.58	1450.64	45	0.12	0.11	0.12	0.91	0.93	8.58	5.57	4.98
		V	-40245.60	80641.20	81061.26	1442.16	45	0.12	0.11	0.12	0.90	0.93	9.10	5.83	5.20
20	500	C	-13359.22	26918.44	27339.90	1071.82	110	0.13	0.12	0.14	0.88	0.90	8.96	5.60	4.95
		V	-13339.79	26879.58	27301.04	1045.94	110	0.13	0.12	0.13	0.88	0.90	10.44	6.51	5.79
	1000	C	-26615.92	53431.83	53922.61	1862.53	110	0.12	0.12	0.13	0.89	0.91	9.80	6.11	5.40
		V	-26592.40	53384.79	53875.57	1820.29	110	0.12	0.12	0.13	0.90	0.91	9.61	6.13	5.45
	1500	C	-40168.73	80537.47	81068.79	3174.26	110	0.13	0.13	0.14	0.87	0.89	9.21	5.92	5.26
		V	-40155.08	80510.17	81041.49	3063.97	110	0.13	0.12	0.13	0.87	0.89	9.62	6.21	5.53
	2000	C	-53508.37	107216.74	107776.83	4607.49	110	0.14	0.13	0.14	0.84	0.86	9.42	6.02	5.32
		V	-53502.81	107205.62	107765.71	4604.81	110	0.14	0.13	0.14	0.84	0.86	9.74	6.21	5.50

Note: C = calibration sample; V = validation sample.

Table 10: Polytomous data: percentages of fit statistics across all replications pointing to the correct model when the true model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
											Singlets	Doublets	Triplets
5	500	C	96	96	95	-	-	-	-	-	100	99	98
		V	96	96	96	-	-	-	-	-	100	98	98
	1000	C	100	100	100	-	-	-	-	-	100	100	100
		V	100	100	100	-	-	-	-	-	100	100	100
	1500	C	99	99	99	-	-	-	-	-	98	98	98
		V	98	98	98	-	-	-	-	-	98	98	98
	2000	C	96	96	96	-	-	-	-	-	100	100	100
		V	98	98	98	-	-	-	-	-	100	100	100
10	500	C	74	71	70	-	-	-	-	-	98	94	94
		V	72	69	66	-	-	-	-	-	98	96	96
	1000	C	70	70	69	-	-	-	-	-	96	94	94
		V	69	69	68	-	-	-	-	-	94	93	93
	1500	C	64	64	63	-	-	-	-	-	97	95	95
		V	62	62	61	-	-	-	-	-	97	95	93
	2000	C	75	75	75	-	-	-	-	-	97	97	97
		V	73	73	72	-	-	-	-	-	97	97	97
15	500	C	79	78	74	76	70	11	66	67	100	100	99
		V	78	78	75	81	71	4	69	70	98	98	98
	1000	C	86	86	85	81	77	8	70	74	100	100	100
		V	85	85	84	82	76	7	71	75	100	100	100
	1500	C	84	83	81	80	74	2	73	76	100	99	99
		V	85	85	85	79	77	4	77	78	100	100	100
	2000	C	82	82	81	79	72	5	67	68	100	100	100
		V	85	85	85	81	74	3	67	69	100	100	100
20	500	C	77	76	71	74	73	11	69	69	100	100	100
		V	79	77	72	76	66	6	65	67	100	100	100
	1000	C	81	81	81	74	70	2	70	70	100	97	97
		V	80	78	75	73	70	5	68	69	98	96	95
	1500	C	88	88	86	78	76	4	72	73	100	100	98
		V	85	84	83	80	77	2	74	74	100	99	99
	2000	C	91	91	90	85	82	3	79	80	100	100	100
		V	93	93	92	82	76	4	72	75	100	100	100

Note: C = calibration sample; V = validation sample.

Table 11: Non-convergence rates by condition when fitting the dominance model to data arising from ideal point and dominance response processes

Dichotomous data								
Data type	2PL				GGUM			
	Simulated data for 15-item scale							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	2%	1%	0%	0%	1.5%	0%	0%	0%
	Simulated data for 30-item scale							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	0%	0%	0%	0%	0.5%	0%	0%	0%
Polytomous data								
Data type	GRM				GGUM			
	Simulated data for 5-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	1.5%	4%	8.5%	1.5%
	Simulated data for 10-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	0%	0.5%	0%	0.5%
	Simulated data for 15-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	1%	0%	0.5%	0%
	Simulated data for 20-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	0%	0%	0%	0%	0%	0%	0%	0%

Note: NC-R = non-convergence rate (i.e., the number of non-converged replications in each condition / the number of total replications in each condition).

Table 12: Non-convergence rates by condition when fitting the ideal point model to data arising from ideal point and dominance response processes

	Dichotomous data							
Data type	2PL				GGUM			
	Simulated data for 15-item scale							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	28.5%	38.5%	56.5%	59.5%	1.5%	1%	2.5%	1.5%
	Simulated data for 30-item scale							
	250	500	1,000	2,000	250	500	1,000	2,000
NC-R	31.5%	52.5%	66%	70%	0.5%	1.5%	0%	0%
	Polytomous data							
Data type	GRM				GGUM			
	Simulated data for 5-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	63%	57%	40%	23.5%	0.5%	0%	0%	1%
	Simulated data for 10-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	67.5%	71%	61.5%	42%	0%	1.5%	2.5%	1.5%
	Simulated data for 15-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	81%	80%	70%	60.5%	1%	1%	3%	6.5%
	Simulated data for 20-item scale							
	500	1,000	1,500	2,000	500	1,000	1,500	2,000
NC-R	90.5%	83%	81.5%	67.5%	1.5%	6.5%	4%	2.5%

Note: NC-R = non-convergence rate (i.e., the number of non-converged replications in each condition / the number of total replications in each condition).

Table 13: Dichotomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the two-parameter logistic model (2PL)

Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
2PL	15	250	C	-2099.28	4258.57	4364.21	90.13	90	0.01	0.00	0.03	1.03	0.96	2.66	1.75	1.49
			V	-2100.83	4261.65	4367.30	91.70	90	0.01	0.00	0.03	1.02	0.95	3.24	1.99	1.66
		500	C	-4274.53	8609.05	8735.49	91.15	90	0.01	0.00	0.02	1.00	0.98	2.64	1.77	1.52
			V	-4261.75	8583.51	8709.94	90.67	90	0.01	0.00	0.02	1.00	0.98	2.83	1.84	1.58
		1000	C	-8515.42	17090.83	17238.06	91.80	90	0.01	0.00	0.02	1.00	0.99	2.70	1.83	1.57
			V	-8515.94	17091.88	17239.12	90.78	90	0.01	0.00	0.02	1.00	0.99	2.85	1.89	1.61
		2000	C	-16954.80	33969.61	34137.64	89.94	90	0.00	0.00	0.01	1.00	1.00	2.81	1.90	1.63
			V	-16946.79	33953.59	34121.61	90.10	90	0.00	0.00	0.01	1.00	0.99	2.87	1.93	1.65
	30	250	C	-4177.90	8475.80	8687.08	406.09	405	0.01	0.00	0.02	1.00	0.98	2.63	1.74	1.48
			V	-4164.35	8448.70	8659.99	415.19	405	0.01	0.00	0.02	0.99	0.98	3.15	1.96	1.64
		500	C	-8413.85	16947.70	17200.57	412.16	405	0.01	0.00	0.02	1.00	0.99	2.70	1.81	1.55
			V	-8418.08	16956.15	17209.03	409.63	405	0.01	0.00	0.02	1.00	0.99	3.00	1.93	1.64
		1000	C	-16862.50	33845.00	34139.46	402.66	405	0.00	0.00	0.01	1.00	1.00	2.71	1.82	1.57
			V	-16871.19	33862.37	34156.84	403.83	405	0.00	0.00	0.01	1.00	1.00	2.87	1.89	1.61
		2000	C	-33621.96	67363.92	67699.97	404.22	405	0.00	0.00	0.01	1.00	1.00	2.72	1.85	1.59
			V	-33589.01	67298.02	67634.07	403.00	405	0.00	0.00	0.01	1.00	1.00	3.05	2.01	1.71
Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
GGUM	15	250	C	-2092.90	4275.80	4434.27	68.75	75	0.01	0.00	0.03	1.01	1.00	0.47	0.28	0.20
			V	-2094.16	4278.32	4436.78	69.07	75	0.01	0.00	0.03	1.01	1.00	2.34	1.31	1.04
		500	C	-4273.30	8636.59	8826.25	68.79	75	0.00	0.00	0.02	1.00	1.00	0.42	0.27	0.21
			V	-4259.29	8608.58	8798.23	68.29	75	0.00	0.00	0.02	1.00	1.00	1.62	1.13	1.00
		1000	C	-8524.59	17139.18	17360.03	70.52	75	0.00	0.00	0.01	1.00	1.00	0.63	0.44	0.37
			V	-8522.54	17135.08	17355.93	68.53	75	0.00	0.00	0.01	1.00	1.00	1.39	1.12	1.05
		2000	C	-16974.61	34039.22	34291.26	69.19	75	0.00	0.00	0.01	1.00	1.00	0.14	0.09	0.07
			V	-16968.22	34026.44	34278.48	68.49	75	0.00	0.00	0.01	1.00	1.00	1.16	1.03	1.00
	30	250	C	-4174.95	8529.90	8846.83	363.96	375	0.00	0.00	0.02	1.00	1.00	3.45	2.23	1.75
			V	-4165.73	8511.45	8828.39	372.43	375	0.01	0.00	0.02	1.00	1.00	21.52	9.71	6.48
		500	C	-8427.32	17034.64	17413.96	368.35	375	0.00	0.00	0.01	1.00	1.00	3.24	2.14	1.76
			V	-8432.93	17045.86	17425.17	366.54	375	0.00	0.00	0.01	1.00	1.00	12.67	6.27	4.47
		1000	C	-16941.85	34063.70	34505.40	368.72	375	0.00	0.00	0.01	1.00	1.00	6.85	4.64	3.89
			V	-16950.25	34080.50	34522.20	370.69	375	0.00	0.00	0.01	1.00	1.00	12.04	7.17	5.69
		2000	C	-33781.33	67742.66	68246.74	390.59	375	0.00	0.00	0.01	1.00	1.00	6.49	4.41	3.70
			V	-33741.09	67662.18	68166.26	380.99	375	0.00	0.00	0.01	1.00	1.00	21.14	12.87	10.59

Note: C = calibration sample; V = validation sample.

Table 14: Dichotomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when the true model was the generalized graded unfolding model (GGUM)

Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
2PL	15	250	C	-2367.84	4795.69	4901.33	146.92	90	0.05	0.03	0.06	0.83	0.85	8.17	3.28	2.13
			V	-2368.92	4797.84	4903.49	139.46	90	0.04	0.03	0.06	0.84	0.86	15.01	7.80	6.04
		500	C	-4743.22	9546.43	9672.87	185.82	90	0.04	0.03	0.05	0.84	0.86	9.55	4.78	3.67
			V	-4742.29	9544.59	9671.02	186.59	90	0.04	0.03	0.05	0.84	0.86	12.45	6.81	5.47
		1000	C	-9549.68	19159.36	19306.60	329.37	90	0.05	0.04	0.06	0.80	0.83	9.89	5.63	4.60
			V	-9543.95	19147.90	19295.13	327.57	90	0.05	0.04	0.06	0.80	0.83	9.67	5.70	4.72
		2000	C	-19041.16	38142.31	38310.34	514.36	90	0.05	0.04	0.05	0.83	0.85	10.24	5.93	4.92
			V	-19048.99	38157.97	38326.00	519.70	90	0.05	0.04	0.05	0.82	0.85	9.23	5.45	4.57
	30	250	C	-4699.35	9518.70	9729.98	705.11	405	0.05	0.05	0.06	0.78	0.80	9.34	3.98	2.66
			V	-4696.81	9513.62	9724.91	703.05	405	0.05	0.05	0.06	0.79	0.80	16.59	8.75	6.78
		500	C	-9402.07	18924.14	19177.01	1049.48	405	0.06	0.05	0.06	0.78	0.79	10.32	5.36	4.14
			V	-9407.90	18935.80	19188.67	1055.26	405	0.06	0.05	0.06	0.77	0.79	13.02	7.24	5.82
		1000	C	-18797.88	37715.77	38010.23	1576.03	405	0.05	0.05	0.06	0.80	0.81	10.87	6.15	5.00
			V	-18793.98	37707.95	38002.42	1597.68	405	0.05	0.05	0.06	0.80	0.81	11.07	6.46	5.32
		2000	C	-37769.22	75658.44	75994.49	2912.20	405	0.05	0.05	0.06	0.77	0.79	11.57	6.79	5.62
			V	-37773.52	75667.03	76003.09	2946.55	405	0.06	0.05	0.06	0.77	0.79	10.16	6.07	5.07
Fitted Model	Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
										Lower	Upper			Singlets	Doublets	Triplets
GGUM	15	250	C	-2345.73	4781.46	4939.92	81.03	75	0.02	0.00	0.04	0.99	0.98	0.32	0.22	0.19
			V	-2346.33	4782.65	4941.12	79.11	75	0.01	0.00	0.04	0.99	0.98	5.30	2.79	2.05
		500	C	-4705.89	9501.78	9691.44	79.02	75	0.01	0.00	0.03	1.00	0.99	0.88	0.76	0.72
			V	-4707.13	9504.25	9693.91	79.27	75	0.01	0.00	0.03	1.00	0.99	3.91	2.37	1.93
		1000	C	-9453.62	18997.24	19218.09	82.56	75	0.01	0.00	0.02	0.99	0.99	0.97	0.90	0.88
			V	-9459.13	19008.27	19229.12	82.42	75	0.01	0.00	0.02	1.00	0.99	2.39	1.66	1.45
		2000	C	-18919.85	37929.70	38181.75	102.12	75	0.01	0.00	0.02	0.99	0.99	1.17	1.08	1.05
			V	-18907.57	37905.14	38157.18	89.30	75	0.01	0.00	0.01	1.00	1.00	2.26	1.71	1.55
	30	250	C	-4609.35	9398.69	9715.63	389.77	375	0.01	0.00	0.02	0.99	0.99	0.61	0.41	0.35
			V	-4608.88	9397.76	9714.70	389.44	375	0.01	0.00	0.03	0.99	0.99	6.08	3.30	2.49
		500	C	-9229.65	18639.29	19018.61	408.79	375	0.01	0.00	0.02	0.99	0.99	1.14	0.93	0.87
			V	-9242.38	18664.76	19044.08	414.69	375	0.01	0.00	0.02	0.99	0.99	4.36	2.68	2.20
		1000	C	-18502.90	37185.79	37627.49	440.07	375	0.01	0.01	0.02	0.99	0.99	1.60	1.33	1.24
			V	-18476.49	37132.99	37574.69	438.61	375	0.01	0.01	0.02	0.99	0.99	4.13	2.83	2.45
		2000	C	-37149.79	74479.58	74983.66	519.88	375	0.01	0.01	0.01	0.99	0.99	1.60	1.36	1.29
			V	-37158.29	74496.57	75000.65	544.24	375	0.01	0.01	0.01	0.99	0.99	3.42	2.50	2.22

Note: C = calibration sample; V = validation sample.

Table 15: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true and the calibrated model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3827.33	7704.66	7810.02	-	-	-	-	-	-	-	2.41	1.98	1.87
		V	-3829.01	7708.01	7813.38	-	-	-	-	-	-	-	5.15	3.91	3.59
	1000	C	-7669.40	15388.80	15511.50	-	-	-	-	-	-	-	2.51	2.02	1.90
		V	-7665.94	15381.87	15504.56	-	-	-	-	-	-	-	5.48	4.03	3.70
	1500	C	-11476.73	23003.47	23136.30	-	-	-	-	-	-	-	2.37	1.97	1.87
		V	-11471.84	22993.69	23126.52	-	-	-	-	-	-	-	5.06	3.90	3.60
	2000	C	-15306.38	30662.77	30802.79	-	-	-	-	-	-	-	2.51	2.00	1.89
		V	-15308.17	30666.35	30806.37	-	-	-	-	-	-	-	5.50	3.99	3.66
10	500	C	-7589.91	15279.83	15490.56	-	-	-	-	-	-	-	2.71	2.04	1.92
		V	-7593.26	15286.50	15497.19	-	-	-	-	-	-	-	6.02	4.07	3.72
	1000	C	-15223.15	30546.31	30791.69	-	-	-	-	-	-	-	2.47	2.01	1.90
		V	-15220.60	30541.19	30786.58	-	-	-	-	-	-	-	5.39	4.05	3.71
	1500	C	-22859.37	45818.75	46084.41	-	-	-	-	-	-	-	2.31	1.94	1.84
		V	-22858.38	45816.77	46082.43	-	-	-	-	-	-	-	4.91	3.81	3.51
	2000	C	-30468.16	61036.33	61316.37	-	-	-	-	-	-	-	2.34	1.97	1.87
		V	-30478.53	61057.07	61337.11	-	-	-	-	-	-	-	5.03	3.91	3.60
15	500	C	-11360.12	22870.24	23186.34	44.58	45	0.01	0.00	0.03	1.05	0.92	2.39	1.97	1.86
		V	-11361.87	22873.72	23189.78	45.73	45	0.01	0.00	0.03	1.02	0.89	5.16	3.90	3.58
	1000	C	-22728.37	45606.73	45974.76	46.35	45	0.01	0.00	0.02	0.99	0.95	2.39	1.98	1.87
		V	-22743.81	45637.63	46005.71	44.66	45	0.01	0.00	0.02	1.01	0.96	5.13	3.92	3.61
	1500	C	-34192.97	68535.94	68934.43	45.70	45	0.01	0.00	0.02	1.00	0.97	2.41	1.99	1.88
		V	-34187.57	68525.14	68923.63	46.65	45	0.01	0.00	0.02	0.99	0.97	5.23	3.97	3.65
	2000	C	-45555.33	91260.66	91680.72	44.97	45	0.00	0.00	0.01	1.00	0.98	2.36	1.95	1.84
		V	-45562.49	91274.98	91695.05	45.16	45	0.00	0.00	0.01	1.00	0.98	5.07	3.83	3.53
20	500	C	-15125.48	30450.96	30872.42	111.08	110	0.01	0.00	0.02	1.00	0.95	2.48	1.98	1.87
		V	-15125.88	30451.74	30873.16	109.43	110	0.01	0.00	0.02	1.02	0.96	5.38	3.92	3.59
	1000	C	-30254.02	60708.04	61198.82	109.00	110	0.00	0.00	0.01	1.00	0.98	2.45	1.99	1.88
		V	-30254.11	60708.21	61198.99	109.23	110	0.00	0.00	0.01	1.00	0.98	5.28	3.95	3.62
	1500	C	-45382.31	90964.61	91495.93	111.13	110	0.00	0.00	0.01	1.00	0.99	2.36	1.96	1.86
		V	-45388.67	90977.33	91508.65	112.24	110	0.00	0.00	0.01	1.00	0.99	5.05	3.88	3.57
	2000	C	-60595.13	121390.25	121950.34	110.00	110	0.00	0.00	0.01	1.00	0.99	2.36	1.97	1.87
		V	-60576.55	121353.10	121913.19	110.81	110	0.00	0.00	0.01	1.00	0.99	5.08	3.90	3.60

Note: C = calibration sample; V = validation sample.

Table 16: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true model was the graded response model (GRM) and the calibrated model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3828.37	7716.74	7843.18	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-3831.73	7723.46	7849.90	-	-	-	-	-	-	-	8.16	5.57	5.07
	1000	C	-7674.31	15408.62	15555.85	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-7670.43	15400.87	15548.10	-	-	-	-	-	-	-	3.97	2.87	2.66
	1500	C	-11485.08	23030.16	23189.56	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-11481.08	23022.16	23181.56	-	-	-	-	-	-	-	3.12	2.32	2.18
	2000	C	-15319.00	30698.00	30866.03	-	-	-	-	-	-	-	0.00	0.00	0.00
		V	-15320.24	30700.49	30868.51	-	-	-	-	-	-	-	2.48	1.94	1.85
10	500	C	-7610.81	15341.61	15594.49	-	-	-	-	-	-	-	2.33	1.83	1.61
		V	-7612.83	15345.64	15598.47	-	-	-	-	-	-	-	9.68	6.88	6.34
	1000	C	-15268.88	30657.76	30952.23	-	-	-	-	-	-	-	2.12	1.76	1.63
		V	-15269.64	30659.29	30953.75	-	-	-	-	-	-	-	6.23	4.73	4.45
	1500	C	-22926.92	45973.84	46292.63	-	-	-	-	-	-	-	1.80	1.52	1.43
		V	-22928.60	45977.19	46295.99	-	-	-	-	-	-	-	4.79	3.80	3.62
	2000	C	-30556.66	61233.33	61569.38	-	-	-	-	-	-	-	1.71	1.45	1.39
		V	-30578.58	61277.16	61613.22	-	-	-	-	-	-	-	4.24	3.44	3.30
15	500	C	-11421.17	23022.34	23401.66	27.01	30	0.01	0.00	0.03	-0.23	0.82	4.72	3.71	3.27
		V	-11421.41	23022.81	23402.08	27.69	30	0.01	0.00	0.03	1.01	0.84	12.23	9.07	8.44
	1000	C	-22860.61	45901.19	46342.84	29.93	30	0.01	0.00	0.02	0.65	0.90	4.94	4.11	3.84
		V	-22873.40	45926.79	46368.49	29.63	30	0.01	0.00	0.02	1.05	0.91	8.70	6.91	6.57
	1500	C	-34399.51	68979.02	69457.21	31.64	30	0.01	0.00	0.02	0.94	0.91	5.12	4.32	4.10
		V	-34361.54	68903.08	69381.27	29.16	30	0.00	0.00	0.02	1.01	0.95	8.18	6.72	6.44
	2000	C	-45796.83	91773.66	92277.74	31.48	30	0.01	0.00	0.01	0.93	0.93	3.83	3.27	3.13
		V	-45805.20	91790.39	92294.47	33.01	30	0.01	0.00	0.02	0.94	0.94	6.34	5.25	5.04
20	500	C	-15219.60	30679.20	31184.95	89.43	90	0.01	0.00	0.02	1.18	0.83	5.33	4.19	3.70
		V	-15231.88	30703.74	31209.45	89.41	90	0.01	0.00	0.02	0.89	0.84	12.82	9.67	9.05
	1000	C	-30455.77	61151.55	61740.48	92.03	90	0.01	0.00	0.02	1.03	0.90	5.20	4.31	4.03
		V	-30475.43	61190.85	61779.78	94.25	90	0.01	0.00	0.02	0.89	0.88	9.19	7.31	6.94
	1500	C	-45710.71	91661.42	92299.01	97.23	90	0.01	0.00	0.01	0.93	0.90	5.77	4.89	4.65
		V	-45699.62	91639.24	92276.83	98.26	90	0.01	0.00	0.01	0.91	0.88	8.82	7.24	6.93
	2000	C	-61033.17	122306.35	122978.46	100.10	90	0.01	0.00	0.01	0.92	0.90	5.93	5.06	4.85
		V	-61026.14	122292.29	122964.40	102.77	90	0.01	0.00	0.01	0.88	0.89	8.67	7.25	6.98

Note: C = calibration sample; V = validation sample.

Table 17: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true and the calibrated model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3356.93	6773.82	6900.18	-	-	-	-	-	-	-	0.57	0.57	0.56
		V	-3350.34	6760.68	6887.12	-	-	-	-	-	-	-	1.61	1.40	1.36
	1000	C	-6766.70	13593.41	13740.64	-	-	-	-	-	-	-	0.75	0.75	0.75
		V	-6760.16	13580.32	13727.56	-	-	-	-	-	-	-	1.24	1.15	1.13
	1500	C	-10176.73	20413.46	20572.85	-	-	-	-	-	-	-	1.04	1.02	1.02
		V	-10171.51	20403.02	20562.42	-	-	-	-	-	-	-	1.37	1.29	1.27
	2000	C	-13549.59	27159.17	27327.20	-	-	-	-	-	-	-	0.88	0.88	0.88
		V	-13565.14	27190.28	27358.31	-	-	-	-	-	-	-	1.15	1.10	1.09
10	500	C	-6694.08	13508.16	13761.04	-	-	-	-	-	-	-	1.31	1.20	1.18
		V	-6686.95	13493.91	13746.78	-	-	-	-	-	-	-	2.25	1.96	1.90
	1000	C	-13263.16	26646.32	26940.79	-	-	-	-	-	-	-	1.71	1.59	1.57
		V	-13233.82	26587.64	26882.10	-	-	-	-	-	-	-	2.54	2.28	2.22
	1500	C	-20078.64	40277.29	40596.08	-	-	-	-	-	-	-	1.85	1.72	1.69
		V	-20097.28	40314.55	40633.35	-	-	-	-	-	-	-	2.17	1.98	1.94
	2000	C	-26417.14	52954.28	53290.34	-	-	-	-	-	-	-	1.48	1.41	1.39
		V	-26389.52	52899.04	53235.10	-	-	-	-	-	-	-	1.72	1.60	1.57
15	500	C	-9733.04	19646.06	20025.33	204.65	30	0.08	0.06	0.09	0.93	0.95	1.15	1.08	1.06
		V	-9763.58	19707.15	20086.47	213.92	30	0.08	0.07	0.10	0.92	0.95	2.46	2.06	2.00
	1000	C	-19490.87	39161.74	39603.44	319.80	30	0.07	0.06	0.09	0.93	0.95	1.31	1.24	1.23
		V	-19507.77	39195.53	39637.23	331.73	30	0.07	0.06	0.09	0.92	0.95	1.75	1.60	1.56
	1500	C	-29255.07	58690.15	59168.34	501.00	30	0.07	0.06	0.08	0.93	0.96	1.46	1.39	1.37
		V	-29280.73	58741.45	59219.64	570.76	30	0.07	0.07	0.08	0.92	0.95	1.75	1.62	1.59
	2000	C	-39006.84	78193.69	78697.77	616.14	30	0.07	0.06	0.08	0.93	0.95	1.42	1.35	1.34
		V	-39016.29	78212.58	78716.66	703.95	30	0.07	0.07	0.08	0.92	0.94	1.65	1.54	1.51
20	500	C	-13032.71	26305.42	26811.17	592.00	90	0.09	0.08	0.10	0.93	0.94	1.65	1.51	1.47
		V	-12994.97	26229.94	26735.69	565.41	90	0.08	0.08	0.09	0.93	0.94	2.47	2.17	2.10
	1000	C	-25972.14	52184.29	52773.22	1100.72	90	0.09	0.08	0.09	0.93	0.94	1.84	1.71	1.68
		V	-25902.65	52045.29	52634.23	1022.04	90	0.08	0.08	0.09	0.93	0.94	2.67	2.36	2.30
	1500	C	-38881.55	78003.09	78640.68	1291.63	90	0.08	0.07	0.08	0.94	0.95	1.76	1.65	1.62
		V	-38856.90	77953.81	78591.40	1340.38	90	0.08	0.07	0.08	0.94	0.95	2.04	1.88	1.84
	2000	C	-51663.29	103566.57	104238.68	1622.75	90	0.07	0.07	0.08	0.94	0.95	1.74	1.64	1.61
		V	-51647.05	103534.11	104206.21	1614.23	90	0.07	0.07	0.08	0.94	0.95	1.96	1.81	1.78

Note: C = calibration sample; V = validation sample.

Table 18: Polytomous data: means of log-likelihood ratios, AIC, BIC, M_2 statistics, df, RMSEA, TLI, CFI, and test-level adjusted χ^2/df ratios across replications when both the true model was the generalized graded unfolding model (GGUM) and the calibrated model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	*df*	RMSEA	RMSEA 90% CI		TLI	CFI	χ^2/df ratios		
									Lower	Upper			Singlets	Doublets	Triplets
5	500	C	-3424.28	6898.53	7003.81	-	-	-	-	-	-	-	6.60	4.21	3.76
		V	-3418.39	6886.79	6992.15	-	-	-	-	-	-	-	8.03	5.19	4.69
	1000	C	-6915.65	13881.30	14003.99	-	-	-	-	-	-	-	6.11	4.16	3.80
		V	-6914.01	13878.02	14000.71	-	-	-	-	-	-	-	7.08	4.75	4.34
	1500	C	-10403.72	20857.44	20990.27	-	-	-	-	-	-	-	7.60	4.89	4.38
		V	-10409.97	20869.94	21002.77	-	-	-	-	-	-	-	8.24	5.22	4.67
	2000	C	-13818.75	27687.51	27827.53	-	-	-	-	-	-	-	6.49	4.35	3.96
		V	-13822.50	27694.99	27835.01	-	-	-	-	-	-	-	7.11	4.71	4.27
10	500	C	-6830.10	13760.21	13970.94	-	-	-	-	-	-	-	7.81	4.97	4.45
		V	-6833.02	13766.05	13976.78	-	-	-	-	-	-	-	9.33	6.01	5.40
	1000	C	-13493.64	27087.29	27332.67	-	-	-	-	-	-	-	8.04	5.10	4.53
		V	-13482.91	27065.82	27311.21	-	-	-	-	-	-	-	9.51	6.01	5.35
	1500	C	-20327.26	40754.51	41020.17	-	-	-	-	-	-	-	7.93	5.16	4.63
		V	-20314.83	40729.66	40995.33	-	-	-	-	-	-	-	8.42	5.45	4.90
	2000	C	-27037.32	54174.63	54454.68	-	-	-	-	-	-	-	8.24	5.21	4.63
		V	-27032.79	54165.58	54445.63	-	-	-	-	-	-	-	8.63	5.46	4.86
15	500	C	-10051.97	20253.92	20569.97	477.44	45	0.13	0.12	0.14	0.87	0.91	8.37	5.30	4.70
		V	-10054.07	20258.13	20574.23	447.38	45	0.12	0.11	0.13	0.88	0.91	11.02	6.73	5.95
	1000	C	-20118.53	40387.05	40755.13	859.69	45	0.13	0.12	0.13	0.88	0.91	8.65	5.53	4.91
		V	-20123.10	40396.20	40764.28	920.23	45	0.13	0.12	0.14	0.88	0.91	9.71	6.15	5.46
	1500	C	-30223.82	60597.65	60996.14	1274.96	45	0.13	0.12	0.13	0.88	0.91	9.23	5.77	5.08
		V	-30222.43	60594.86	60993.35	1285.04	45	0.13	0.12	0.13	0.88	0.91	9.82	6.12	5.39
	2000	C	-40242.75	80635.51	81055.58	1450.64	45	0.12	0.11	0.12	0.91	0.93	8.58	5.57	4.98
		V	-40245.60	80641.20	81061.26	1442.16	45	0.12	0.11	0.12	0.90	0.93	9.10	5.83	5.20
20	500	C	-13359.45	26918.89	27340.35	1076.26	110	0.13	0.12	0.14	0.88	0.90	8.83	5.55	4.90
		V	-13340.32	26880.64	27302.10	1053.34	110	0.13	0.12	0.13	0.88	0.90	10.32	6.46	5.74
	1000	C	-26629.44	53458.87	53949.65	1860.65	110	0.12	0.12	0.13	0.89	0.91	8.80	5.63	5.00
		V	-26592.09	53384.17	53874.95	1828.26	110	0.12	0.12	0.13	0.90	0.91	10.79	6.64	5.88
	1500	C	-40167.89	80535.78	81067.10	3175.40	110	0.13	0.13	0.14	0.87	0.89	9.19	5.91	5.25
		V	-40153.90	80507.80	81039.12	3059.24	110	0.13	0.12	0.13	0.87	0.89	9.61	6.20	5.52
	2000	C	-53511.88	107223.77	107783.86	4622.04	110	0.14	0.13	0.14	0.84	0.86	9.41	6.01	5.32
		V	-53501.79	107203.59	107763.68	4607.33	110	0.14	0.13	0.14	0.84	0.86	9.72	6.20	5.50

Note: C = calibration sample; V = validation sample.

Table 19: Dichotomous data: percentages of fit statistics across all replications pointing to the correct model

True Model	Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
												Singlets	Doublets	Triplets
2PL	15	250	C	6	92	100	0	57	44	49	53	5	4	2
			V	12	92	100	0	59	43	51	55	23	15	12
		500	C	32	97	100	0	51	39	45	47	4	4	3
			V	36	97	100	1	52	39	43	48	5	4	3
		1000	C	69	100	100	0	52	39	49	49	1	1	0
			V	71	100	100	1	55	44	41	47	1	1	1
		2000	C	80	100	100	0	54	37	44	47	1	1	1
			V	85	100	100	1	61	51	49	52	0	0	0
	30	250	C	38	100	100	1	50	46	40	41	21	20	20
			V	34	100	100	0	53	41	44	48	100	99	93
		500	C	61	100	100	0	55	45	45	49	24	24	24
			V	65	100	100	0	52	43	40	45	100	92	74
		1000	C	98	100	100	6	66	53	49	51	30	30	30
			V	95	100	100	7	62	47	50	54	98	84	74
		2000	C	93	100	100	10	76	58	57	60	29	29	28
			V	94	100	100	11	70	54	55	57	72	64	58
GGUM	15	250	C	81	66	24	98	93	7	96	96	100	98	95
			V	82	67	26	98	92	10	95	95	96	95	95
		500	C	79	72	46	100	99	1	100	100	99	99	99
			V	83	73	49	99	98	2	99	99	96	94	93
		1000	C	82	78	68	98	96	4	98	99	100	100	100
			V	85	82	71	99	99	0	99	100	98	97	97
		2000	C	78	75	72	99	98	2	99	99	100	100	100
			V	74	73	67	97	97	1	97	97	95	92	92
	30	250	C	94	81	53	100	100	0	100	100	100	99	99
			V	96	86	60	100	100	0	100	100	98	98	96
		500	C	94	92	76	99	99	0	99	99	100	100	100
			V	95	88	77	99	99	0	99	99	97	96	96
		1000	C	88	84	78	100	100	0	100	100	100	100	100
			V	88	85	81	100	100	0	100	100	94	92	92
		2000	C	91	91	89	100	99	1	99	99	100	100	100
			V	92	92	88	100	100	0	100	100	95	92	91

Note: C = calibration sample; V = validation sample.

Table 20: Polytomous data: percentages of fit statistics across all replications pointing to the correct model when the true model was the graded response model (GRM)

Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
											Singlets	Doublets	Triplets
5	500	C	66	91	100	-	-	-	-	-	0	0	0
		V	70	97	100	-	-	-	-	-	79	65	62
	1000	C	79	99	100	-	-	-	-	-	0	0	0
		V	86	99	100	-	-	-	-	-	27	25	27
	1500	C	93	100	100	-	-	-	-	-	0	0	0
		V	93	100	100	-	-	-	-	-	14	12	11
	2000	C	98	100	100	-	-	-	-	-	0	0	0
		V	98	100	100	-	-	-	-	-	4	3	4
10	500	C	86	100	100	-	-	-	-	-	25	22	19
		V	92	99	100	-	-	-	-	-	84	77	77
	1000	C	98	100	100	-	-	-	-	-	21	21	21
		V	100	100	100	-	-	-	-	-	45	45	46
	1500	C	100	100	100	-	-	-	-	-	19	19	18
		V	100	100	100	-	-	-	-	-	30	31	33
	2000	C	100	100	100	-	-	-	-	-	17	17	17
		V	100	100	100	-	-	-	-	-	17	17	18
15	500	C	98	100	100	1	62	53	48	66	50	47	46
		V	97	100	100	4	67	61	36	27	96	93	92
	1000	C	99	100	100	8	71	59	42	70	57	56	55
		V	100	100	100	9	61	52	40	29	67	66	67
	1500	C	100	100	100	8	58	43	33	60	55	55	55
		V	100	100	100	11	69	57	50	41	58	59	61
	2000	C	100	100	100	11	75	59	53	77	50	51	51
		V	100	100	100	14	65	50	38	30	47	50	52
20	500	C	100	100	100	7	65	56	43	76	63	60	58
		V	99	100	100	8	68	64	40	38	95	89	89
	1000	C	100	100	100	18	74	64	46	79	67	68	65
		V	100	100	100	16	74	67	47	45	79	81	81
	1500	C	100	100	100	21	72	58	53	78	67	67	67
		V	100	100	100	21	70	60	51	49	75	77	78
	2000	C	100	100	100	27	75	58	57	81	73	73	73
		V	100	100	100	28	77	62	52	51	70	74	74

Note: C = calibration sample; V = validation sample.

Table 21: Polytomous data: percentages of fit statistics across all replications pointing to the correct model when the true model was the generalized graded unfolding model (GGUM)

Items	N	Type	LL	AIC	BIC	M_2	RMSEA	RMSEA Overlap	TLI	CFI	χ^2/df ratios		
											Singlets	Doublets	Triplets
5	500	C	96	96	95	-	-	-	-	-	100	99	98
		V	96	96	96	-	-	-	-	-	100	98	98
	1000	C	100	100	100	-	-	-	-	-	100	100	100
		V	100	100	100	-	-	-	-	-	100	100	100
	1500	C	99	99	99	-	-	-	-	-	98	98	98
		V	98	98	98	-	-	-	-	-	98	98	98
	2000	C	96	96	96	-	-	-	-	-	100	100	100
		V	98	98	98	-	-	-	-	-	100	100	100
10	500	C	74	71	70	-	-	-	-	-	98	94	94
		V	72	69	66	-	-	-	-	-	98	96	96
	1000	C	70	70	69	-	-	-	-	-	96	94	94
		V	70	70	69	-	-	-	-	-	95	94	94
	1500	C	64	64	63	-	-	-	-	-	97	95	95
		V	62	62	61	-	-	-	-	-	97	95	93
	2000	C	75	75	75	-	-	-	-	-	97	97	97
		V	73	73	72	-	-	-	-	-	97	97	97
15	500	C	79	78	74	76	70	11	66	67	100	100	99
		V	78	78	75	81	71	4	69	70	100	100	100
	1000	C	85	85	84	80	76	8	69	73	100	100	100
		V	85	85	84	82	76	7	71	75	100	100	100
	1500	C	84	83	81	80	74	2	73	76	100	99	99
		V	85	85	85	79	77	4	77	78	100	100	100
	2000	C	82	82	81	79	72	5	67	68	100	100	100
		V	85	85	85	80	74	3	67	69	100	100	100
20	500	C	77	76	71	74	73	11	69	69	100	100	100
		V	79	78	72	76	67	6	66	68	100	100	100
	1000	C	81	81	81	74	70	2	70	70	100	97	97
		V	80	78	75	73	70	5	68	69	100	98	97
	1500	C	87	87	86	78	76	4	72	73	100	100	98
		V	85	84	83	80	77	2	74	74	100	99	99
	2000	C	91	91	90	85	82	3	79	80	100	100	100
		V	93	93	92	82	76	4	72	75	100	100	100

Note: C = calibration sample; V = validation sample.

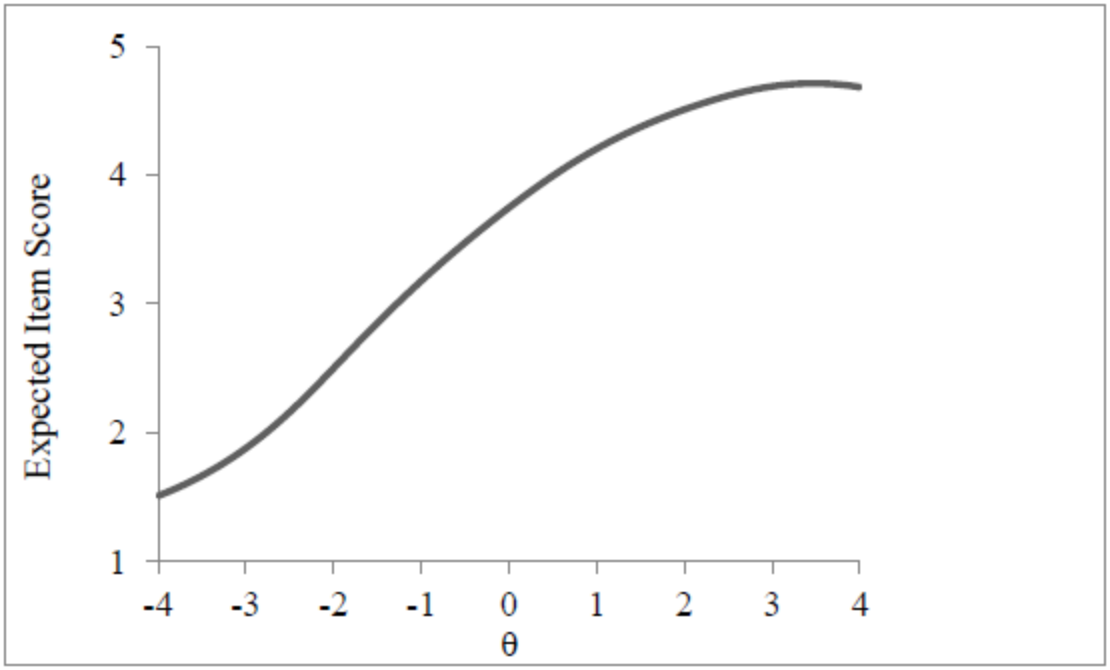

Figure 1. Example of the Dominance Response Process

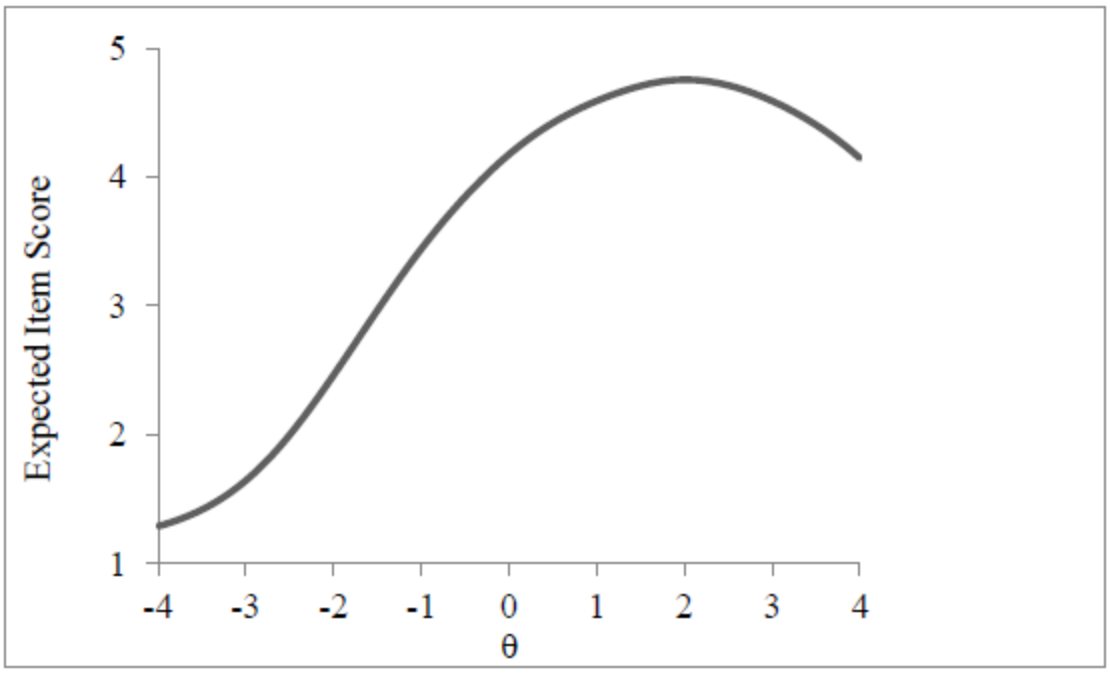

Figure 2. Example of the Ideal Point Response Process