International Journal of Testing

ISSN: 1530-5058 (Print) 1532-7574 (Online) Journal homepage: https://www.tandfonline.com/loi/hijt20

Detecting Curvilinear Relationships: A Comparison of Scoring Approaches Based on Different Item Response Models

Mengyang Cao, Q. Chelsea Song & Louis Tay

To cite this article: Mengyang Cao, Q. Chelsea Song & Louis Tay (2018) Detecting Curvilinear Relationships: A Comparison of Scoring Approaches Based on Different Item Response Models, International Journal of Testing, 18:2, 178-205, DOI: 10.1080/15305058.2017.1345913

To link to this article: https://doi.org/10.1080/15305058.2017.1345913

Published online: 07 Nov 2017.

International Journal of Testing, 18: 178–205, 2018

ISSN: 1530-5058 print / 1532-7574 online
DOI: 10.1080/15305058.2017.1345913

Routledge
Taylor & Francis Group

Detecting Curvilinear Relationships: A Comparison of Scoring Approaches Based on Different Item Response Models

Mengyang Cao and Q. Chelsea Song
University of Illinois at Urbana-Champaign

Louis Tay
Purdue University

There is a growing use of noncognitive assessments around the world, and recent research has posited an ideal point response process underlying such measures. A critical issue is whether the typical use of dominance approaches (e.g., average scores, factor analysis, and the Samejima's graded response model) in scoring such measures is adequate. This study examined the performance of an ideal point scoring approach (e.g., the generalized graded unfolding model) as compared to the typical dominance scoring approaches in detecting curvilinear relationships between scored trait and external variable. Simulation results showed that when data followed the ideal point model, the ideal point approach generally exhibited more power and provided more accurate estimates of curvilinear effects than the dominance approaches. No substantial difference was found between ideal point and dominance scoring approaches in terms of Type I error rate and bias across different sample sizes and scale lengths, although skewness in the distribution of trait and external variable can potentially reduce statistical power. For dominance data, the ideal point scoring approach exhibited convergence problems in most conditions and failed to perform as well as the dominance scoring approaches. Practical implications for scoring responses to Likert-type surveys to examine curvilinear effects are discussed.

Keywords: Curvilinear relationship, ideal point model, dominance model, item response theory

Within the United States and across the world, there is increased interest in the use of noncognitive measures of assessments (Elosua & Iliescu, 2012; Ramesh,

Mengyang Cao is now at Facebook.
Correspondence should be sent to Mengyang Cao, 1 Hacker Way, Menlo Park, CA 94025, USA. E-mail: mcao@fb.com

Hazucha, & Bank, 2008; Smith, Gorske, Wiggins, & Little, 2010). Despite the prevalence of scoring approaches to noncognitive measures based on the dominance response model (e.g., average scores, factor analysis, and the Samejima's graded response model), recent research has suggested that the ideal point model be considered as an alternative item response model in scoring surveys of individual differences such as attitudes, personality, interests, and affect (e.g., Chernyshenko, Stark, Drasgow, & Roberts, 2007; Cho, Drasgow, & Cao, 2015; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009; Tay & Kuykendall, 2016). Unlike the dominance model, which assumes that the probability of endorsing an item monotonically increases as the latent trait standing increases (e.g., logistic item response theory models), the ideal point model assumes that the probability of endorsing an item is proportional to the distance between the latent trait standing and the location of the item (Coombs, 1964), thus more accurately describes the responses to noncognitive individual differences surveys where individuals compare an item to his or her "ideal point" when responding (Drasgow et al., 2010; Tay & Drasgow, 2012).

While there have been several advantages found in using the appropriate item response model for noncognitive individual differences surveys (e.g., Tay & Drasgow, 2012), one aspect that recently received attention is the improvement in detecting the curvilinear relationships between predictors and outcomes. Specifically, Carter and colleagues (2014) empirically compared the performance of four different scoring approaches in detecting the curvilinear relationships between conscientiousness and job performance, and found that when conscientiousness measures were scored by an ideal point approach, they more consistently exhibited significant curvilinear links to job performance than when scored by dominance approaches. This is an intriguing finding and demonstrates the importance of using the appropriate item response model for scoring, and has important implications for assessing curvilinear effects in educational and psychological research (e.g., Bowman, 2013; Grant, 2013; Knifsend & Graham, 2012). This could potentially impact selection outcomes should there not be a linear relationship between noncognitive predictors and outcomes. For instance, it may be possible that individuals who are overly extraverted may be overly socially dominant or talkative and have worse interpersonal outcomes. A critical issue is whether dominance approaches to scoring noncognitive predictors occlude these curvilinear relations.

Nevertheless, past research has shown that scoring personality data (including conscientiousness) with either an ideal point or a dominance model did not result in substantially different scores; that is, ideal point and dominance scores tended to show very high correlations (although divergences occurred among higher scores) (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; Stark et al., 2006). As such, it is important to use a simulation study to further examine whether ideal point scoring as compared to dominance scoring leads to

substantially better detection of curvilinear effects. First, it is difficult to ascertain with empirical data whether the more consistently significant results from an ideal point scoring approach was a consequence of higher power, or was due to a higher Type I error rate. Without knowing the "true" effect size of the curvilinear relationships, we are also unable to assess the bias and estimation accuracy of different scoring approaches. Second, although similar results were obtained from three different samples (Carter et al., 2014), it is not known whether these results would replicate in other samples with different trait distributions and survey design parameters (e.g., scale length and sample size). Third, findings in Carter and colleagues (2014) were based on the fact that the ideal point model showed better fit than the dominance model to personality data. It is unclear whether the ideal point model would still demonstrate superior performance in detecting curvilinear relationships if the responses are generated based on the dominance model.

To examine these issues, we conducted Monte Carlo simulations to investigate how different scoring approaches based on either the dominance or the ideal point response model affect the detection of curvilinear relationships. Specifically, we compared the performance of these scoring approaches in terms of Type I error rate, power, bias, and estimation accuracy. We also explored whether the performance would be affected by survey design factors (e.g., scale length and sample size), data generation methods (dominance vs. ideal point), and the distribution of variables (e.g., skewness).

Item Response Models and Scoring

Conventionally constructed individual difference measures were assumed to follow the dominance response model, and were scored using the dominance-based scoring approaches (Drasgow et al., 2010). The dominance model assumes a monotonic relationship between responses to a survey and the latent trait levels, which is consistent with the assumptions in classical test theory (CTT) (Allen & Yen, 1979). Due to the simplicity in model assumptions, dominance scoring approaches have been widely adopted in the field of educational and psychological measurement. For example, the CTT scoring approach of summing or averaging the endorsed responses across all items on the scale also relies on the assumption of dominance response model—that responses to any item are monotonically associated with true scores (Chernyshenko et al., 2007). The dominance model is also the underlying assumption of traditional factor analysis, which assumes a monotonic (and linear) relationship between responses and factor scores (Harris, 1967). Most item response theory (IRT) models also rely on the dominance assumption (Stark et al., 2006). For example, both the two-parameter logistic model (2PLM) for dichotomous responses and the Samejima's graded response model (SGRM) (Samejima, 1969) for Likert-type

responses assume a nonlinear but monotonically increasing relationship between observed responses and the latent trait (i.e., θ).

On the contrary, the ideal point model assumes that the probability of endorsing an item depends on how closely an individual's latent trait matches the location of the item (Coombs, 1964). Such assumption resembles the response process underlying noncognitive individual difference assessments, where respondents often compare the content described by the item (i.e., the item location) with their own attributes, and endorse the item only if traits and locations are close (Tay et al., 2009; Tay & Drasgow, 2012); that is, an individual is likely to reject an item when his or her trait standing is either much higher than the item location (i.e., rejecting from above), or much lower than the item location (i.e., rejecting from below). Empirical results consistently demonstrate that the ideal point model exhibits better fit than the dominance model to noncognitive measures such as attitudes, personality, vocational interests, emotional intelligence, and emotions (Cho et al., 2015; Roberts, Laughlin, & Wedell, 1999; Stark et al., 2006; Tay et al., 2009; Tay & Drasgow, 2012; Tay & Kuykendall, 2016). As such, dominance scoring approaches based on the monotonicity assumptions may be less appropriate for noncognitive measures (Drasgow et al., 2010).

Despite the fundamental differences in model assumptions between dominance and ideal point scoring approaches, empirical research has found that the correlations between scores estimated by the dominance model and those estimated by the ideal point model were close to unity (Chernyshenko et al., 2007; Weekers & Meijer, 2008). Moreover, there was only negligible difference between dominance and ideal point scoring approaches in predictive validity, indicated by the linear correlation between the estimated predictor scores and values of the external variable (Cao, Drasgow, & Cho, 2015). However, such results do not necessarily suggest that dominance and ideal point scoring approaches would generate same results in all situations. Both simulation and empirical results showed that applying the dominance model to score ideal point data would incorrectly place a proportion of individuals with high attributes to the middle of the trait continuum (Roberts et al., 1999). Although such a subtle change in rank ordering may not be reflected when detecting linear relationships, it may significantly impact results when detecting curvilinear relationships, which requires more accurate estimates of the predictor, especially for individuals with extremely high latent trait levels (Carter et al., 2014). To demonstrate this, we simulated a sample of 500 with a $N(0,1)$ distribution, generated responses to a 15-item scale based on an ideal point model, and estimated the trait levels using both dominance and ideal point models. As illustrated in Figure 1, trait estimates by the two models are very close in the middle range of the distribution, but exhibit discrepancies on the high-end and low-end of the distribution. Specifically, the dominance model tends to underestimate trait levels on the high-end compared to the ideal point model.

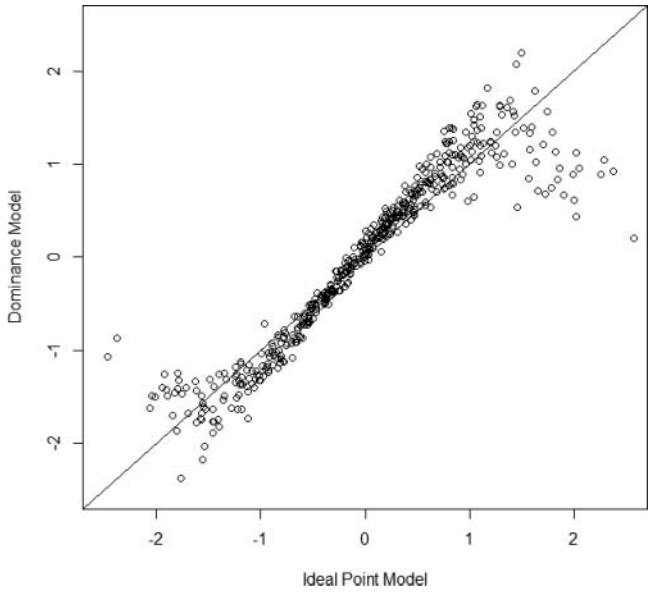

FIGURE 1
A demonstration of the relationship between trait estimates by the ideal point model and by the dominance model.

Detecting Curvilinear Relationships

Although most research in the field of education and psychology has been focused on investigating linear relationships among variables because of its simplicity in modeling and interpretation, linearity may not appropriately describe all relationships (Nikolaeva, Bhatnagar, & Ghose, 2015). Indeed, curvilinear relationships, among a variety of nonlinear forms of associations have been commonly found in psychological literature (Johnson, 2014). A curvilinear effect can be a U-shaped curve that describes a resurgence after a certain period of decline, or an inverted U-shaped curve that suggests a too-much-of-a-good-thing effect (Pierce & Aguinis, 2013), For example, a U-shaped curvilinear relationship was found between conscientiousness and counterproductive work behaviors (CWB) (opposite of an inverted-U shape because of the undesirable outcome, but reflecting a too-much-of-a-good-thing effect), such that extremely conscientious people may engage in more deviant behaviors in a company than individuals with medium levels of conscientiousness (Carter et al., 2014). The relationship between statistics anxiety and performance among college students was an inverted U-shaped curve, suggesting that there was an optimal level of anxiety that was most productive for students (Keeley, Zayac, & Correia, 2008).

A commonly used approach to detecting curvilinear relationships is polynomial regression. Specifically, the outcome variable *Y* is regressed on both the linear and the quadratic forms of the predictor variable *X*, as expressed in the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon.$$

To examine the curvilinear effect, one tests the null hypothesis $\beta_2 = 0$ by comparing $\hat{\beta}_2 / SE(\hat{\beta}_2)$ to a critical value obtained from the *t*-distribution. A significant $\hat{\beta}_2$. would suggest the existence of a curvilinear relationship.

As demonstrated in Figure 1, when responses are generated based on the ideal point model, the dominance scoring approach would mistakenly place individuals with high attributes on the middle range of the distribution. Such misplacement might lead to failures in detecting the curvilinear relationship between the dominance scores and the criterion. On the other hand, when individuals use a dominance response process, mixed results have been found in previous research that whether the ideal point model should still be used as the scing approach. It has been proposed that the ideal point model is more flexible than the dominance mel such that the parameterization of the ideal point model allows it to adequately fit dominance data as well, while a previous simulation study showed that this is only true when the ideal point model was applied to dichotomous dominance data (Tay, Ali, Drasgow, & Williams, 2011). It also remains unclear whether the flexibility of the ideal point model would affect the detection of curvilinear relationships of dominance responses.

Apart from item response models, the detection of curvilinear relationships can be affected by several other factors. First, previous research suggests that the measurement errors of the predictor can severely influence the detection of curvilinear relationships (Johnson, 2014). Specifically, if the observed responses are treated as true scores without accounting for measurement errors, the magnitude of quadratic effects can be reduced (Agustin & Singh, 2005; Busemeyer & Jones, 1983). Thus, accurately estimating the predictor scores is critical in detecting curvilinear relationships. Latent trait scoring methods such as factor analysis (FA) and IRT should be considered, as they can account for measurement errors. In this simulation study, we will compare the performance of average scores based on observed responses with the performance of latent trait scoring methods. Besides scoring methods, the length of the measurement scale can also affect measurement errors, as longer scales normally provide more information and consequently more accurate estimates of the latent trait. We will manipulate scale length as a factor to examine how it affects the detection of curvilinear relationships.

Second, the skewness in predictor and outcome may also affect the detection of curvilinear effects. On the predictor side, most IRT scoring practice adopts Bayesian methods, which almost always assume a normal distribution as the priori for latent trait (Hambleton, Swaminathan, & Rogers, 1991; Harris, 1967). However, such assumption can be easily violated on different occasions. For example, when personality assessments are administered in high-stakes situations such as personnel selection scenarios, some individuals might be motived to inflate their scores, resulting in a negatively skewed distribution of personality trait (Viswesvaran & Ones, 1999). Such skewed distribution of the predictor may affect the accuracy of the trait estimates, which can lead to lower power in detecting curvilinear relationships. On the outcome side, normality is also an essential assumption for hypothesis testing in regression models. Nonetheless, many outcome variables are often skewed. For example, it has been shown in a variety of settings that job performance is skewed (O'Boyle & Aguinis, 2012). A skewed outcome may potentially violate the assumption of regression models, and consequently yielding an inaccurate detection of curvilinear effects.

Third, as with all significance testing, power to detect effects can be enhanced by a larger sample size (Cohen, 1994). Therefore, curvilinear effects would be more frequently detected with a larger sample size. This is another factor that also needs to be considered in seeking to understand when curvilinear effects may be detected depending on the type of response model used.

In the present study, we focus on addressing a practical question: When examining the curvilinear relationship between scored trait (i.e., the *predictor*) and external variable (i.e., the *outcome*), how will the measurement model adopted to score the predictor influence the detection of the curvilinear relationship? Since true trait values are unknown in practical situations, researchers may be unaware of the level of measurement inaccuracy of the trait estimates obtained from different measurement models, which may obstruct the detection of curvilinear relationships. Based on the aforementioned discussion, our simulation examines several factors including: (a) data generation method (dominance vs. ideal point); b) survey design factors (scale length and sample size); and (c) skewness in predictor and outcome.

METHOD

Study Design

We conducted Monte Carlo simulations to compare the detection of curvilinear effects in different conditions. Specifically, we manipulated the following variables:

1. Sample size: (a) 250, (b) 500, (c) 1000, or (d) 2000.

2. Scale length: (a) 15 items or (b) 30 items.
3. Response data generation: (a) ideal point model or (b) dominance model
4. Coefficient of the quadratic effect: (a) 0 (no quadratic effect) or (b) 0.1. Note that the magnitude of quadratic effect was chosen to be consistent with the empirical findings in Carter and colleagues (2014), where most of the coefficients of the quadratic terms were found to be around 0.1.
5. Skewness in predictor: (a) 0 (not skewed) or (b) –0.5 (negatively skewed).
6. Skewness in outcome: (a) 0 (not skewed) or (b) –0.5 (negatively skewed).

Altogether, we had 128 conditions, and each condition was replicated 200 times. For each replication, item parameters, person parameters, responses, and outcome variables were randomly generated, and results were aggregated across 200 replications to obtain stable results representing the corresponding condition.

Latent Trait and External Variable Generation

For the nonskewed trait and outcome condition, the latent trait values (i.e., θ) were sampled from a $N(0,1)$ distribution. The outcome variable values were then simulated using the following equation:

$$Y_i = \beta_1 \theta_i + \beta_1 \theta_i^2 + e_i$$

where Y_i denotes the outcome value for the ith simulee; e_i is the random error sampled from a $N(0,1)$ distribution; the linear effect β_1 was set to 0.2 to be consistent with the empirical findings in Carter and colleagues (2014); β_2 was set to either 0.1 or 0 to indicate the quadratic effect.

For skewed predictor or outcome conditions, skewness was simulated using the power method (Fleishman, 1978), which has been widely adopted in studies involving simulating skewness (e.g., Curran, West, & Finch, 1996; Enders, 2001; Nasser & Wisenbaker, 2003). In particular, the method utilizes the following power function to transform a normal distribution to a skewed distribution, while maintaining the means and standard deviations invariant:

$$Y = a + bX + cX^2 + dX^3$$

where X indicates the normally distributed variable, and Y is the skewed variable after transformation. The desired degree of skewness can be achieved by using the combination of coefficients provided in Table 1 of Fleishman (1978).

In this study, we chose skewness = −0.5 to represent a moderately skewed non-normal distribution. The skewness of predictor was simulated based on the N (0,1) distribution, and the skewness of outcome was simulated by transforming the outcome variable with the power method.

Response Generation

The generalized graded unfolding model (GGUM) (Roberts, Donoghue, & Laughlin, 2000) is arguably the most frequently used ideal point IRT model. The probability for respondent j to endorse option k of item i given by the GGUM can be expressed as

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

where α, δ, and τ represent the discrimination, location, and subjective threshold parameters, respectively.

We used the polytomous GGUM (Roberts et al., 2000) to generate the ideal point data. Item parameters were generated based on the procedures proposed by Roberts and colleagues (2002), which has been widely used in previous simulation studies (e.g., Carter & Zickar, 2011; Tay, Ali, Drasgow, & Williams, 2011; Wang, Tay, & Drasgow, 2013). Specifically, discrimination parameters were generated from a uniform (0.5, 2) distribution; location parameters were sampled from equally spaced intervals from –2 to 2, and the middle 4 parameters were dropped because responses generated from those items could not be estimated by the SGRM (Tay et al., 2011); subjective threshold parameters were generated from a uniform (–1.4, –0.4) distribution, with successive values generated with the following equation:

$$\tau_{ik-1} = \tau_{ik} - 0.25 + e_{ik-1},$$

where $e_{ik\text{-}1}$ indicates random errors sampled from a $N(0, 0.04)$ distribution. Responses were then generated on a 5-point scale by comparing a random number from the $U(0, 1)$ distribution to the cumulative distribution function of the GGUM equation with the simulated item parameters and latent trait values.

Dominance responses were generated on a 5-point scale with the SGRM (Samejima, 1969). We followed the procedures in previous simulation studies (e.g., Reise & Waller, 1990; Tay et al., 2011) to sample item parameters and generate response data. In particular, we sampled the a-parameters from a

uniform (0.5, 0.8) distribution, and four b_k-parameters respectively from four uniform distributions: (−2.0, −1.0), (−1.0, 0.0), (0.0, 1.0), and (1.0, 2.0).

Model Fitting and Scoring

Consistent with Carter and colleagues (2014), we used four different scoring approaches to obtain the estimated scores of the predictor. First, we computed the average responses across all items as the classical test theory (CTT) scores. For ideal point responses, items with negative location parameters were reverse-coded before computing the CTT scores. The average scores were then standardized before computing the quadratic forms to reduce multicollinearity (Ganzach, 1997). Second, we obtained the factor analysis (FA) scores by extracting a one-factor principle axis solution and computing the regression-based factors scores. Third, we used the SGRM as a dominance IRT scoring approach. Specifically, we estimated SGRM item parameters based on the simulated responses using the MULTILOG 7.0 software with default settings (Thissen, 2003), then estimated the person scores as the SGRM scores using the maximum a posteriori (MAP) scoring approach provided by the MULTILOG 7.0 software. Note that all the above three scoring methods were based on the dominance item response model, which assumed a monotonic relationship between latent trait scores and observed responses. Fourth, we estimated GGUM item and person parameters with the GGUM2004 software (Roberts, Fang, Cui, & Wang, 2006), which utilizes an expected a posteriori (EAP) estimation approach. The GGUM scoring method is based on the ideal point response model assumption. Besides the four scoring approaches, we also examined the quadratic effects of the simulated thetas as the baseline model.

Detecting the Curvilinear Relationship

The following polynomial regression model was fitted to the predictor scores estimated by each of the scoring methods to detect the curvilinear relationship:

$$Y_i = \beta_0 + \beta_1 \hat{\theta}_i + \beta_2 \hat{\theta}_i^2 + e_i$$

where $\hat{\theta}_i$ denotes the estimated score of the predictor. We examined five indices to assess the performance of each scoring method.

Type I Error Rate. For conditions with no quadratic effects, we recorded the ratio out of 200 replications with significant quadratic terms as the Type I error rate. We sought to determine whether obtained Type I error rates were well controlled at the nominal rate of 0.05.

TABLE 1
Type I Error Rate for Different Simulation Conditions

Response data		Ideal point data								Dominance data							
Scale length		15 items				30 items				15 items				30 items			
Sample size		250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
No skewness	Baseline	0.06	0.04	0.06	0.05	0.06	0.09	0.06	0.06	0.03	0.04	0.03	0.02	0.07	0.04	0.06	0.05
	CTT	0.06	0.06	0.04	0.09	0.04	0.04	0.06	0.05	0.08	0.06	0.04	0.04	0.05	0.04	0.07	0.07
	FA	0.08	0.05	0.04	0.07	0.05	0.05	0.07	0.05	0.07	0.05	0.03	0.04	0.05	0.04	0.07	0.07
	SGRM	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.06	0.08	0.04	0.04	0.03	0.06	0.05	0.06	0.07
	GGUM	0.05	0.03	0.06	0.05	0.07	0.08	0.05	0.06	0.08	0.08	0.05	0.05	—	—	—	—
Skewed predictor only	Baseline	0.08	0.04	0.08	0.05	0.06	0.04	0.04	0.04	0.09	0.05	0.04	0.06	0.06	0.04	0.04	0.02
	CTT	0.07	0.07	0.05	0.04	0.07	0.03	0.07	0.07	0.08	0.06	0.08	0.08	0.05	0.06	0.07	0.05
	FA	0.06	0.05	0.07	0.08	0.07	0.04	0.05	0.08	0.08	0.07	0.07	0.08	0.06	0.06	0.07	0.06
	SGRM	0.06	0.05	0.06	0.08	0.07	0.04	0.04	0.09	0.07	0.07	0.05	0.07	0.05	0.05	0.04	0.05
	GGUM	0.05	0.06	0.10	0.09	0.05	0.05	0.06	0.03	*0.15*	*0.00*	*0.20*	*0.15*	—	—	—	—
Skewed outcome only	Baseline	0.06	0.05	0.08	0.06	0.05	0.02	0.06	0.03	0.07	0.04	0.05	0.05	0.05	0.05	0.03	0.03
	CTT	0.04	0.06	0.05	0.05	0.09	0.05	0.07	0.08	0.06	0.03	0.05	0.06	0.06	0.05	0.06	0.08
	FA	0.06	0.05	0.06	0.05	0.09	0.05	0.08	0.07	0.05	0.02	0.05	0.06	0.06	0.05	0.06	0.08
	SGRM	0.05	0.05	0.06	0.06	0.08	0.04	0.09	0.07	0.05	0.03	0.05	0.05	0.06	0.05	0.05	0.07
	GGUM	0.04	0.04	0.06	0.08	0.05	0.03	0.06	0.05	*0.00*	*0.15*	*0.00*	*0.05*	—	—	—	—
Both skewed predictor and outcome	Baseline	0.06	0.06	0.09	0.06	0.07	0.04	0.06	0.04	0.03	0.04	0.05	0.05	0.05	0.06	0.07	0.08
	CTT	0.04	0.06	0.06	0.06	0.05	0.04	0.03	0.07	0.08	0.07	0.10	0.17	0.06	0.05	0.12	0.13
	FA	0.06	0.06	0.07	0.08	0.07	0.05	0.03	0.07	0.08	0.07	0.11	0.18	0.05	0.06	0.11	0.13
	SGRM	0.05	0.05	0.07	0.08	0.06	0.05	0.03	0.08	0.09	0.07	0.09	0.13	0.06	0.06	0.08	0.11
	GGUM	0.04	0.07	0.13	0.12	0.08	0.04	0.07	0.09	*0.00*	*0.10*	*0.05*	*0.25*	—	—	—	—

Note. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

Power. For conditions where quadratic effects were simulated, we defined power as the ratio out of 200 replications where the estimated coefficient of the quadratic term (i.e., $\hat{\beta}_2$) was found to be significant at a nominal Type I error rate of $\alpha = 0.05$.

Bias. For conditions with simulated quadratic effects, we computed the difference between the estimated coefficient of the quadratic term and the simulated effect size (i.e., 0.1) for each replication, and averaged across all replications as the indicator of estimation bias.

Root-Mean-Square Error (RMSE). For conditions with simulated quadratic effects, we also computed the square root of the average of squared difference between the estimated coefficient of the quadratic term and the simulated effect size (i.e., 0.1) across all replications. The RMSE represents the accuracy of the estimation of the curvilinear effect.

Change in R^2. For conditions with simulated quadratic effects, we obtained the R^2 statistics for both the model with only the linear term and the model with both linear and quadratic terms, and computed the differences between the R^2 statistics of the two models. The change in R^2 indicates the amount of additional variance explained by the curvilinear effect.

RESULTS

Simulation results are summarized in Tables 1–3. When dominance data were simulated, the GGUM program encountered convergence problems in most replications when skewed predictors were simulated, and failed to converge when scale length = 30. Therefore, we were only able to run 20 replications for most conditions when scale length = 15, except that the results in "no skewness" conditions were still based on 200 replications.

Type I Error Rate

Simulation results on Type I error rate are presented in Table 1. In general, most scoring approaches exhibited satisfactory control over Type I error rates regardless of sample size, scale length, and skewness. For ideal point data, only two conditions reported Type I error rate over 0.10: the GGUM approach where ideal point data were simulated on a 15-item scale and a sample size of 500 and 1000, with both predictor and outcome simulated as skewed. For dominance data, higher Type I error rates were found in conditions where skewness was simulated and the GGUM was used for scoring. When skewness was simulated in

both predictor and outcome, Type I error rates were found to be above 0.10 even when dominance approaches were used. Overall, there were no consistently discernable differences among scoring approaches.

Power

Simulation results of statistical power are presented in Table 2. When data were generated based on the ideal point model, the GGUM approach exhibited substantially higher power than the dominance scoring approaches in almost all conditions, and had similar performance to the baseline model. Although dominance scoring approaches tended to have lower power compared to the ideal point approach, the difference among the three dominance scoring approaches was relatively small and showed variations across conditions.

As expected, the power increased for all scoring approaches as the sample size increased. When sample size was small (i.e., $n = 250$ or $n = 500$), the ideal point scoring approach showed much higher power than the dominance scoring approaches. Specifically, the power of the GGUM approach ranged from 0.29 to 0.85, whereas the power of dominance approaches ranged from 0.11 to 0.46. However, this difference in power between ideal point and dominance scoring approaches diminished as sample size increased. At $n = 2000$, the GGUM approach can almost always detect the curvilinear relationship, whereas the dominance approaches also consistently showed good power (above 0.80).

For scale length, there seemed to be a consistent effect for the GGUM approach, such that the power increased when more items were included in the scale. The effect of scale length for dominance approaches was mixed and appeared to be dependent on the sample size. For example, more items would generally grant the dominance approaches more power at $n = 250$, 500 or 2000, but would reduce the power on most occasions when $n = 1000$. However, such conclusion did not apply to all conditions.

The power of all the scoring approaches also appeared to be heavily influenced by the skewness in the predictor and the outcome. For the GGUM approach, power always decreased when the sample size was small ($n = 250$ or $n = 500$), regardless of where skewness occurred. When the sample size was medium ($n = 1000$), power was less affected by skewness in the predictor or the outcome. Skewness did not affect power when sample size was large ($n = 2000$). For dominance approaches, however, skewness did not appear to negatively impact the power to detect curvilinear effects. Nevertheless, it may be because the power to detect curvilinear effects was substantially lower (i.e., floor effects) than the ideal point scoring approach.

When data were generated based on the dominance model, dominance scoring approaches were generally more powerful than the ideal point

TABLE 2
Power for Different Simulation Conditions

Response data		Ideal point data								Dominance data							
Scale length		15 items				30 items				15 items				30 items			
Sample size		250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
No skewness	Baseline	0.52	0.87	0.99	1.00	0.61	0.91	1.00	1.00	0.56	0.82	0.99	1.00	0.62	0.86	1.00	1.00
	CTT	0.13	0.27	0.56	0.82	0.18	0.28	0.54	0.78	0.42	0.73	0.96	1.00	0.44	0.76	0.99	1.00
	FA	0.18	0.33	0.59	0.86	0.24	0.28	0.70	0.92	0.42	0.72	0.96	1.00	0.45	0.79	0.98	1.00
	SGRM	0.18	0.33	0.59	0.86	0.24	0.37	0.72	0.92	0.25	0.72	0.96	1.00	0.45	0.78	0.99	1.00
	GGUM	0.52	0.82	0.97	1.00	0.58	0.85	10.00	1.00	0.15	0.43	0.85	0.99	—	—	—	—
Skewed predictor only	Baseline	0.50	0.79	0.98	1.00	0.49	0.81	0.98	1.00	0.47	0.71	0.97	1.00	0.46	0.81	0.96	1.00
	CTT	0.14	0.30	0.62	0.84	0.16	0.32	0.58	0.91	0.25	0.42	0.83	0.97	0.31	0.60	0.85	0.99
	FA	0.15	0.32	0.65	0.89	0.17	0.32	0.59	0.92	0.26	0.41	0.86	0.97	0.29	0.57	0.86	0.99
	SGRM	0.15	0.30	0.62	0.82	0.17	0.33	0.60	0.93	0.27	0.44	0.88	0.99	0.33	0.68	0.91	1.00
	GGUM	0.40	0.67	0.92	1.00	0.42	0.77	0.97	1.00	*0.25*	*0.30*	*0.65*	*0.85*	—	—	—	—
Skewed outcome only	Baseline	0.44	0.81	0.98	1.00	0.40	0.73	0.97	1.00	0.41	0.68	0.97	1.00	0.53	0.81	0.97	1.00
	CTT	0.11	0.29	0.58	0.84	0.17	0.33	0.47	0.84	0.30	0.50	0.90	1.00	0.43	0.74	0.94	1.00
	FA	0.18	0.34	0.67	0.87	0.23	0.46	0.64	0.93	0.29	0.49	0.89	1.00	0.46	0.75	0.94	1.00
	SGRM	0.14	0.33	0.62	0.83	0.21	0.43	0.61	0.92	0.28	0.51	0.92	1.00	0.47	0.75	0.94	1.00
	GGUM	0.29	0.73	0.96	1.00	0.40	0.67	0.96	1.00	*0.15*	*0.45*	*0.90*	*0.94*	—	—	—	—
Both skewed predictor and outcome	Baseline	0.54	0.79	0.97	1.00	0.49	0.79	0.99	1.00	0.40	0.80	0.99	10.00	0.50	0.81	0.96	1.00
	CTT	0.20	0.28	0.55	0.79	0.18	0.33	0.54	0.86	0.18	0.50	0.75	0.95	0.34	0.55	0.81	0.99
	FA	0.22	0.32	0.61	0.82	0.19	0.34	0.57	0.87	0.19	0.48	0.77	0.96	0.32	0.56	0.80	0.99
	SGRM	0.16	0.29	0.57	0.74	0.22	0.35	0.62	0.91	0.22	0.52	0.83	0.99	0.34	0.65	0.91	1.00
	GGUM	0.41	0.66	0.89	1.00	0.42	0.70	0.97	1.00	*0.15*	*0.45*	*0.50*	*0.90*	—	—	—	—

Note. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

approach in detecting curvilinear relationships, though such difference diminished as sample size increased. Note that in many conditions only 20 replications were performed for the ideal point scoring approach, thus the results should be interpreted with caution. Similar to the results with the ideal point data, the power of all scoring approaches was lower in conditions where skewness was simulated, especially for GGUM scoring when sample size was large. In those conditions, the power of dominance scoring approaches was higher for scale length = 30.

Estimation Accuracy (RMSE)

Table 3 summarizes the simulation results on RMSE across all conditions. When responses were generated based on the ideal point model, the GGUM approach provided more accurate estimation (i.e., lower RMSE) of the quadratic coefficient compared to the dominance approaches, and in many conditions the accuracy of the GGUM approach was close to the baseline. There was no consistently noticeable difference among the dominance scoring approaches. Sample size consistently affected the estimation accuracy for all scoring approaches, such that larger sample size was always associated with lower RMSE. Neither scale length nor skewness exhibited systematic impact on estimation accuracy. When data were generated based on the dominance model, the GGUM approach exhibited much lower accuracy than dominance approaches when sample size was small, but were almost as accurate as dominance approaches when sample size was medium or large.

Bias and Change in R^2

For both ideal point and dominance, most scoring approaches generated trivial negative bias (i.e., most values between –0.02 to 0) in the estimation of quadratic terms. Overall, no systematic impact of sample size, scale length, skewness, or scoring method was found on estimation bias. Similarly, the R^2 changes due to quadratic terms are also trivial (i.e., most values from 0 to 0.02) in most conditions. There is also no noticeable difference between simulation conditions. Values of bias and R^2 change of all conditions are presented in the Appendix.

Additional Simulation Conditions

To further explore the impact of scale lengths and quadratic effect size on curvilinear relationship detection, we examined several additional simulation conditions. Considering that many inventories in social sciences have fewer than 15 items, we performed simulations for additional conditions with scale

TABLE 3
Root-Mean-Squared Error for Different Simulation Conditions

Response data		Ideal point data								Dominance data							
Scale length		15 items				30 items				15 items				30 items			
Sample size		250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
No skewness	Baseline	0.049	0.030	0.024	0.016	0.047	0.029	0.024	0.017	0.050	0.033	0.023	0.016	0.049	0.034	0.024	0.016
	CTT	0.080	0.059	0.043	0.035	0.083	0.063	0.042	0.037	0.057	0.040	0.027	0.019	0.061	0.044	0.027	0.020
	FA	0.074	0.062	0.044	0.031	0.082	0.059	0.040	0.029	0.064	0.045	0.032	0.025	0.068	0.051	0.032	0.027
	SGRM	0.074	0.057	0.042	0.033	0.073	0.053	0.038	0.029	0.091	0.039	0.025	0.017	0.057	0.039	0.024	0.019
	GGUM	0.054	0.040	0.028	0.018	0.053	0.038	0.029	0.018	0.122	0.064	0.040	0.025	-	-	-	-
Skewed predictor only	Baseline	0.048	0.034	0.026	0.016	0.049	0.035	0.026	0.017	0.049	0.040	0.026	0.017	0.048	0.034	0.026	0.019
	CTT	0.081	0.060	0.042	0.034	0.081	0.058	0.043	0.028	0.060	0.048	0.036	0.031	0.057	0.040	0.031	0.020
	FA	0.085	0.064	0.043	0.031	0.083	0.059	0.043	0.029	0.066	0.049	0.036	0.026	0.060	0.043	0.031	0.020
	SGRM	0.077	0.060	0.042	0.035	0.075	0.054	0.042	0.027	0.061	0.047	0.033	0.025	0.055	0.038	0.029	0.020
	GGUM	0.059	0.041	0.031	0.020	0.055	0.043	0.030	0.020	*0.077*	*0.076*	*0.033*	*0.031*	-	-	-	-
Skewed outcome only	Baseline	0.047	0.030	0.025	0.018	0.049	0.034	0.028	0.026	0.051	0.037	0.024	0.019	0.044	0.032	0.025	0.019
	CTT	0.091	0.060	0.042	0.035	0.086	0.059	0.048	0.036	0.058	0.046	0.031	0.024	0.048	0.034	0.027	0.020
	FA	0.084	0.062	0.042	0.031	0.088	0.056	0.040	0.027	0.064	0.047	0.032	0.021	0.054	0.037	0.028	0.020
	SGRM	0.087	0.058	0.041	0.035	0.081	0.053	0.042	0.031	0.057	0.045	0.028	0.020	0.046	0.034	0.026	0.020
	GGUM	0.050	0.034	0.026	0.019	0.052	0.036	0.027	0.025	*0.089*	*0.061*	*0.025*	*0.024*	-	-	-	-
Both skewed predictor and outcome	Baseline	0.052	0.033	0.025	0.017	0.049	0.034	0.023	0.017	0.049	0.033	0.022	0.018	0.043	0.033	0.024	0.017
	CTT	0.077	0.061	0.044	0.038	0.087	0.064	0.041	0.031	0.065	0.046	0.037	0.028	0.052	0.046	0.032	0.024
	FA	0.083	0.063	0.043	0.034	0.089	0.065	0.040	0.030	0.069	0.047	0.033	0.027	0.056	0.047	0.031	0.022
	SGRM	0.072	0.060	0.045	0.040	0.082	0.061	0.038	0.030	0.062	0.043	0.032	0.027	0.049	0.042	0.011	0.022
	GGUM	0.056	0.040	0.031	0.023	0.057	0.041	0.025	0.018	*0.098*	*0.051*	*0.043*	*0.030*	-	-	-	-

Note. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

length = 8[1]. As shown in Table 4, results are quite similar to conditions where scale length equals 15 or 30. Type I error rates were well controlled for all scoring approaches. Although the power was slightly lower than conditions with longer scale length, results still consistently showed that the GGUM exhibited higher power than dominance approaches for ideal point data, whereas the dominance scoring approaches had high power than the GGUM for dominance data. Results on bias, RMSE, and change in R^2 are also similar to scale length = 15 or 30 (details available upon request).

As shown in Table 2, in conditions where sample size is large, the power is close to unity, indicating that the quadratic effect size may have reached the ceiling effect. To address this issue, we simulated additional conditions with quadratic coefficients ranging from 0.02 to 0.08.[2] As shown in Table 5, although power decreases as the magnitude of quadratic term decreases, results are consistent with quadratic effect size = 0.10 in that scoring approaches almost always exhibit higher power when applied to the corresponding data. Results on bias, RMSE, and change in R^2 are available upon request.

DISCUSSION

With the development of ideal point models (e.g., Roberts et al., 2000), more attention has been paid on the use of the appropriate measurement models for noncognitive individual difference surveys (Drasgow et al., 2010). While research has shown that the correlation between dominance scoring and ideal point scoring is very high, recent empirical research suggests that differences in scoring may affect the detection of curvilinear effects (Carter et al., 2014). However, it is not known if this is a result of Type I error or the ability to detect "true" curvilinear effects. To examine this issue, and the different factors that may impact the detection of curvilinear effects, we conducted a Monte Carlo simulation varying survey design factors (i.e., sample size, scale length) and variable distributions (i.e., skewness).

Our results revealed that when response data followed an ideal point model, the ideal point scoring approach performed the best in detecting curvilinear effects compared to dominance approaches, and this was not a result of false positive (i.e., Type I error). In fact, the ideal point scoring approach had similar performance compared to the baseline situation. This

[1]We thank an anonymous reviewer for this suggestion.
[2]We thank an anonymous reviewer for this suggestion.

TABLE 4
Type I Error Rate and Power for 8-item Conditions

Type I Error Rate								
Response data	Ideal point data				Dominance data			
Sample size	250	500	1000	2000	250	500	1000	2000
Baseline	0.04	0.05	0.04	0.04	0.05	0.04	0.04	0.05
CTT	0.02	0.05	0.02	0.06	0.05	0.04	0.04	0.03
FA	0.00	0.07	0.04	0.10	0.06	0.04	0.04	0.04
SGRM	0.00	0.08	0.02	0.10	0.06	0.03	0.04	0.05
GGUM	0.08	0.05	0.10	0.10	*0.00*	*0.00*	*0.00*	*0.00*
				Power				
Baseline	0.61	0.89	0.99	1.00	0.61	0.86	0.99	1.00
CTT	0.15	0.27	0.47	0.62	0.40	0.64	0.89	0.99
FA	0.23	0.38	0.60	0.68	0.38	0.63	0.88	0.99
SGRM	0.11	0.24	0.47	0.52	0.38	0.64	0.88	1.00
GGUM	0.44	0.68	0.94	0.99	*0.30*	*0.35*	*0.80*	*1.00*

Note. Number of items = 8. No skewness on predictor and outcome. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

TABLE 5
Power for Quadratic Coefficient = 0.02, 0.04, 0.06, 0.08 Conditions

Response data	Ideal point data															
Quadratic Coefficient	0.02				0.04				0.06				0.08			
Sample size	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
Baseline	0.08	0.14	0.14	0.25	0.18	0.36	0.38	0.60	0.26	0.42	0.78	0.32	0.38	0.58	0.64	0.78
CTT	0.06	0.14	0.12	0.10	0.08	0.12	0.08	0.16	0.08	0.10	0.12	0.14	0.08	0.08	0.12	0.15
FA	0.06	0.10	0.10	0.20	0.12	0.18	0.20	0.34	0.14	0.14	0.36	0.22	0.08	0.14	0.26	0.22
SGRM	0.05	0.10	0.10	0.20	0.08	0.10	0.14	0.14	0.12	0.12	0.26	0.16	0.10	0.14	0.24	0.30
GGUM	0.07	0.20	0.12	0.25	0.08	0.32	0.36	0.50	0.20	0.46	0.68	0.32	0.28	0.68	0.68	0.72
								Dominance data								
Baseline	0.07	0.08	0.15	0.24	0.14	0.18	0.43	0.74	0.31	0.43	0.78	0.94	0.44	0.70	0.96	1.00
CTT	0.07	0.06	0.11	0.14	0.10	0.16	0.26	0.53	0.19	0.30	0.50	0.83	0.28	0.47	0.82	0.98
FA	0.07	0.07	0.09	0.13	0.11	0.15	0.27	0.53	0.19	0.30	0.52	0.83	0.31	0.47	0.84	0.98
SGRM	0.07	0.07	0.11	0.16	0.12	0.17	0.28	0.53	0.19	0.33	0.58	0.83	0.31	0.52	0.82	0.98
GGUM	*0.05*	*0.20*	*0.10*	*0.15*	*0.10*	*0.15*	*0.25*	*0.70*	*0.10*	*0.40*	*0.45*	*0.90*	*0.05*	*0.30*	*0.85*	*1.00*

Note. Number of items = 15. No skewness on predictor and outcome. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

indicates that the ideal point scoring approach, which allows for curvilinear relation between trait and observed scores, result in latent trait values that well fit to actual responses. On the other hand, the dominance scoring approaches only achieved adequate power when sample size reached 2000, a sample size that is not common in personnel selection situations. In addition, we also found that the ideal point scoring approach generally had lower bias and estimation accuracy based on the RMSE than dominance scoring approaches.

When responses were generated by a dominance model, we found that the ideal point scoring approach encountered convergence problems in many conditions. Compared to dominance scoring approaches, the ideal point scoring approach also exhibited lower power, more bias, and less accuracy estimation of curvilinear effects. These results suggested that the ideal point model is not necessarily more flexible than the dominance model. If a misspecified model is used to score the predictor, researchers are less likely to discover the curvilinear relationship between the predictor and the external variable.

When considering the impact of different factors (i.e., sample size, scale length, skewness in predictor and/or outcome) on the performance of ideal point approach and dominance approaches, we found that sample size was influential for statistical power, such that a larger sample size corresponded to a higher power for both the ideal point and dominance scoring approaches. Importantly, when the sample size is small, an ideal point scoring approach tends to have substantially higher power in detecting curvilinear relationship compared to dominance scoring approaches. However, this difference diminished when sample size reached 2000. Sample size did not appear to have large effects on the difference between ideal point scoring and dominance scoring approaches in terms of Type I error rate, bias or RMSE. Similarly, scale length and skewness did not have large effects on the difference between the ideal point scoring approach and dominance scoring approaches in terms of power, Type I error rate, bias and RMSE. One issue that we found was that skewness appeared to lower the power for ideal point scoring as compared to dominance scoring approaches, but is likely due to floor effects of power in dominance scoring approaches, where dominance scoring approaches have substantially lower power.

Theoretical and Practical Implications

The findings of the current Monte Carlo simulation study are consistent with the empirical study by Carter and colleagues (2014) such that when response data fit the ideal point model, the ideal point scoring approach exhibit better performance in detecting curvilinear relationships compared to dominance scoring approaches. Contrary to the findings in Carter and colleagues (2014) where

significant curvilinear relationships were found 100% of the time when ideal point scoring was adopted, we found that the power of detecting depends on several factors, including sample size, scale lengths, and response distribution. We also extended the findings of Carter and colleagues (2014) by demonstrating that when responses are generated based on the dominance model, the ideal point model should not be used for scoring.

Our results suggest that to maximize the power to detect curvilinear effects for Likert-type survey data, it is important to first conduct a model comparison to determine which response model better fits the data (Tay et al., 2011). As shown in our simulation results, choosing a misspecified model to score responses would significantly reduce the power of detecting curvilinear relationships, as well as the accuracy in estimating the curvilinear coefficients. We believe this is largely because a misspecified model generally leads to worse model fit to the responses than the correct measurement model (Tay et al., 2011). Figures 2a and 2b demonstrate the fit plots for ideal point responses to a simulated item estimated by the SGRM and GGUM, respectively. The GGUM, which is the theoretically correct model in this case, shows almost perfect fit to the data, whereas the SGRM exhibits large discrepancies between the theoretical empirical response curves at low- and high-ends of the trait continuum. This supports our earlier argument that the dominance model fails to provide accurate estimation of low or high trait levels for ideal point responses (see Figure 1), as the model exhibits poor fit to the data. Similarly, Figures 3a and 3b also demonstrate that the GGUM exhibits poor fit to dominance responses at low- and high-ends of the trait continuum.

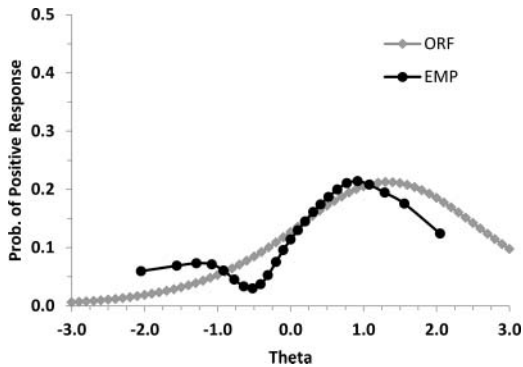

FIGURE 2A
Fit plot for responses to Option 3 of an example item simulated from the ideal point model, and estimated by Samejima's graded response model (SGRM). ORF = Option response function; EMP = Empirical response function.

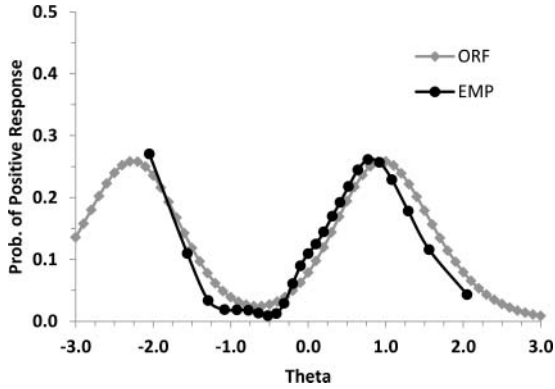

FIGURE 2B
Fit plot for responses to Option 3 of an example item simulated from the ideal point model, and estimated by the generalized graded unfolding model (GGUM). ORF = Option response function; EMP = Empirical response function.

Limitations and Future Directions

Although the results of the current study have important theoretical and practical implications, our study has some drawbacks that require future research. First, we considered only simulation conditions of data generation, sample size, scale length, and skewness. It is possible that other factors such as location of inflection point (e.g., Converse & Oswald, 2014) and scale format (e.g., pairwise comparison; Stark, Chernyshenko, Drasgow & White, 2012) may also affect the detection of curvilinear relationship. For example, Converse and Oswald (2014) found that in personnel selection scenarios, a misspecification of selection procedures in terms of linearity and inflection point may result in substantially negative effects on the mean performance of those selected. Second, the software available for dominance and ideal point score estimation adopts different estimation methods (MAP vs. EAP), which may potentially influence the estimation accuracy of the two models. To minimize the impact of different estimation approaches, we adopted similar settings for both programs, including fixing the number of expectation-maximization (EM) cycles at 200 and the number of theta grid points at 30. Third, we found that when responses were simulated based on the dominance model, the ideal point scoring approach encountered severe convergence problems in many conditions. It is desirable to acquire empirical data and examine if the nonconvergence is due to the simulated data or the nature of the ideal point model fitting dominance data.

Future studies can also investigate how well the ideal point scoring approach performs in the detection of moderator effects in curvilinear relations

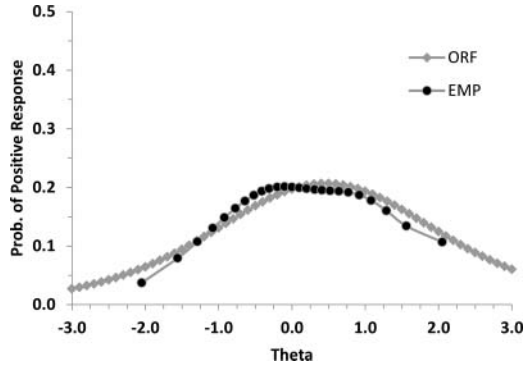

FIGURE 3A
Fit plot for responses to Option 3 of an example item simulated from the dominance model, and estimated by Samejima's graded response model (SGRM). ORF = Option response function; EMP = Empirical response function.

(e.g., Janssen, 2001; Le et al., 2011). A common practice in detecting moderator of curvilinear relation is to look at both the interaction between (1) potential moderator and the linear term of predictor-outcome relation and (2) potential moderator and the quadratic term of predictor-outcome relation (Le et al., 2011). Considering the potential multicollinearity issue introduced by including the interaction term, it needs to be substantiated with future simulation research whether the ideal point scoring can also facilitate the detection of curvilinear relationships between an interaction predictor and outcomes.

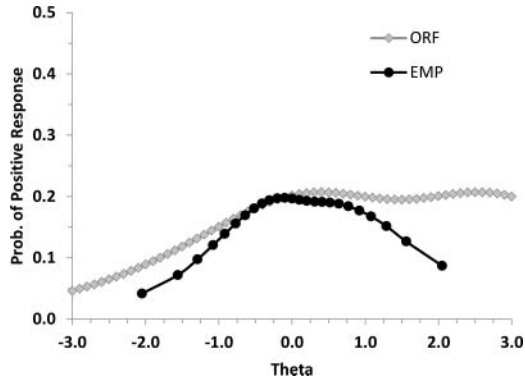

FIGURE 3B
Fit plot for responses to Option 3 of an example item simulated from the dominance model, and estimated by the generalized graded unfolding model (GGUM). ORF = Option response function; EMP = Empirical response function.

CONCLUSIONS

Based on Monte Carlo simulation results, we found that when responses to surveys followed the ideal point model, the ideal point scoring approach produced much higher power in detecting curvilinear relationships than dominance scoring approaches, while still maintaining well controlled Type I error rates. Other factors such as sample size, scale length, and skewness in predictor and outcome also affected the performance of different scoring approaches in detecting curvilinear relationships. When dominance data are simulated, the ideal point scoring approach was found to exhibit less power and accuracy than the dominance scoring approaches.

REFERENCES

Agustin, C., & Singh, J. (2005). Curvilinear effects of consumer loyalty determinants in relational exchanges. *Journal of Marketing Research*, *42*(1), 96–108. doi:10.1509/jmkr.42.1.96.56961

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Bowman, N. A. (2013). How much diversity is enough? The curvilinear relationship between college diversity interactions and first-year student outcomes. *Research in Higher Education*, *54*(8), 874–894. doi:10.1007/s11162-013-9300-0

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*(3), 549–562. doi:10.1037/0033-2909.93.3.549

Cao, M., Drasgow, F., & Cho, S. (2015). Developing ideal intermediate personality items for the ideal point model. *Organizational Research Methods*, *18*(2), 25–275. doi:1094428114555993

Carter, N. T., & Zickar, M. J. (2011). A comparison of the LR and DFIT frameworks of differential functioning applied to the generalized graded unfolding model. *Applied Psychological Measurement*, *35*(8), 623–642. doi:10.1177/0146621611427898

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

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. doi:10.1207/S15327906MBR3604_03

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. doi:10.1037/1040-3590.19.1.88

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological Assessment*. Advance online publication. doi:10.1037/pas0000132

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. doi:10.1037/0003-066X.49.12.997

Converse, P. D., & Oswald, F. L. (2014). Thinking ahead: Assuming linear versus nonlinear personality-criterion relationships in personnel selection. *Human Performance*, *27*(1), 61–79. doi:10.1080/08959285.2013.854367

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

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. doi:10.1037/1082-989X.1.1.16

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *3*(4), 465–476. doi:10.1111/j.1754-9434.2010.01273.x

Elosua, P., & Iliescu, D. (2012). Tests in Europe: Where we are and where we should go. *International Journal of Testing*, *12*, 157–175. doi: 10.1080/15305058.2012.657316

Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, *6*(4), 352–370. doi:10.1037/1082-989X.6.4.352

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521–532. doi:10.1007/BF02293811

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*(3), 235–247.

Grant, A. M. (2013). Rethinking the extraverted sales ideal: The ambivert advantage. *Psychological Science*, *24*(6), 1024–1030. doi:10.1177/0956797612463706

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*, Thousand Oaks, CA: Sage.

Harris, C. W. (1967). On factors and factor scores. *Psychometrika*, *32*(4), 363–379.

Janssen, O. (2001). Fairness perceptions as a moderator in the curvilinear relationships between job demands, and job performance and job dissatisfaction. *Academy of Management Journal*, *44*(5), 1039–1050. doi:10.2307/3069447

Johnson, J. S. (2014). Nonlinear analyses in sales research: theoretical bases and analytical considerations for polynomial models. *Journal of Personal Selling & Sales Management*, *34*(4), 302–317. doi:10.1080/08853134.2014.903804

Keeley, J., Zayac, R., & Correia, C. (2008). Curvilinear relationships between statistics anxiety and performance among undergraduate students: Evidence for optimal anxiety. *Statistics Education Research Journal*, *7*(1), 4–15.

Knifsend, C. A., & Graham, S. (2012). Too much of a good thing? How breadth of extracurricular participation relates to school-related affect and academic outcomes during adolescence. *Journal of Youth and Adolescence*, *41*(3), 379–389. doi:10.1007/s10964-011-9737-4

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

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, *63*(5), 729–757. doi:10.1177/0013164403258228

Nikolaeva, R., Bhatnagar, A., & Ghose, S. (2015). Exploring curvilinearity through fractional polynomials in management research. *Organizational Research Methods*. Advance online publication. doi:1094428115584006.

O'Boyle, E., Jr., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, *65*(1), 79–119. doi:10.1111/j.1744-6570.2011.01239.x

Pierce, J. R., & Aguinis, H. (2013). The too-much-of-a-good-thing effect in management. *Journal of Management*, *39*, 313–338. doi:10.1177/0149206311410060

Ramesh, A., Hazucha, J. F., & Bank, J. (2008). Using personality data to make decisions about global managers. *International Journal of Testing*, *8*, 346–366. doi:10.1080/15305050802435110

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement*, *14*, 45–58. doi:10.1177/014662169001400105

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement*, *59*, 211–233. doi:10.1177/00131649921969811

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.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3–32. doi:10.1177/01466216000241001

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A windows-based program to estimate parameters in the generalized graded unfolding model. *Applied Psychological Measurement*, *30*(1), 64–65. doi:10.1177/0146621605280141

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4), 100. doi:10.1002/j.2333-8504.1968.tb00153.x

Smith, S. R., Gorske, T. T., Wiggins, C., & Little, J. A. (2010). Personality assessment use by clinical neuropsychologists. *International Journal of Testing*, *10*, 6–20. doi:10.1080/15305050903534787

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*, 25–39. doi:10.1037/0021-9010.91.1.25

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, *15*(3), 463–487. doi:1094428112444611

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*, 280–295. doi:10.1177/0146621610390674

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

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. doi:10.1177/1094428112439709

Tay, L., & Kuykendall, L. (2016). Why self-reports of happiness and sadness may not necessarily contradict bipolarity: A psychometric review and proposal. *Emotion Review*, *9*(2), 146–154. doi:10.1177/1754073916637656

Thissen, D. (2003). *Multilog 7: Multiple categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, *59*(2), 197–210. doi:10.1177/00131649921969802

Wang, W., Tay, L., & Drasgow, F. (2013). Detecting differential item functioning of polytomous items for an ideal point response process. *Applied Psychological Measurement*, *37*(4), 316–335. doi:10.1177/0146621613476156

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a Dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment*, *24*, 65–77. doi:10.1027/1015-5759.24.1.65

APPENDIX

TABLE A
Bias for Different Simulation Conditions

Response data		Ideal point data								Dominance data							
Scale length		15 items				30 items				15 items				30 items			
Sample size		250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
No skewness	Baseline	0.003	−0.004	0.001	−0.001	0.004	0.001	0.001	0.000	−0.003	−0.006	−0.003	0.000	0.003	0.000	−0.001	0.000
	CTT	−0.040	−0.019	−0.019	−0.021	−0.020	−0.027	−0.021	−0.025	0.001	−0.003	0.002	0.004	0.004	0.008	0.004	0.006
	FA	−0.007	−0.006	−0.005	−0.007	0.002	−0.027	−0.001	−0.004	0.012	0.008	0.012	0.014	0.014	0.019	0.015	0.016
	SGRM	−0.025	−0.016	−0.018	−0.021	−0.009	−0.015	−0.010	−0.014	−0.049	−0.006	−0.001	0.001	0.001	0.005	0.001	0.003
	GGUM	0.006	0.008	0.007	0.002	0.007	0.013	0.009	0.004	−0.061	−0.021	−0.003	0.007	–	–	–	–
Skewed predictor only	Baseline	−0.003	−0.005	−0.002	0.000	−0.004	−0.002	0.000	0.001	0.001	−0.001	0.001	0.000	−0.004	0.002	0.000	0.001
	CTT	−0.028	−0.018	−0.015	−0.017	−0.016	−0.015	−0.013	−0.007	−0.024	−0.171	−0.019	−0.023	−0.017	−0.009	−0.012	−0.010
	FA	−0.013	−0.005	−0.001	−0.002	−0.007	−0.008	−0.005	0.001	−0.007	−0.012	−0.002	−0.007	−0.008	0.000	−0.003	−0.001
	SGRM	−0.030	−0.019	−0.019	−0.022	−0.019	−0.018	−0.015	−0.011	−0.012	−0.017	−0.007	−0.011	−0.013	−0.005	−0.009	−0.007
	GGUM	−0.007	−0.002	−0.007	−0.006	−0.004	0.008	0.003	0.002	*−0.019*	*−0.024*	*−0.019*	*−0.019*	–	–	–	–
Skewed outcome only	Baseline	−0.021	−0.014	−0.013	−0.012	−0.022	−0.020	−0.021	−0.021	−0.011	−0.015	−0.011	−0.012	−0.006	−0.010	−0.012	−0.012
	CTT	−0.057	−0.022	−0.018	−0.022	−0.031	−0.023	−0.032	−0.027	−0.023	−0.024	−0.015	−0.184	−0.002	−0.006	−0.008	−0.009
	FA	−0.027	−0.010	−0.005	−0.009	−0.013	−0.006	−0.013	−0.010	−0.004	−0.008	0.002	0.000	0.008	0.003	0.001	0.001
	SGRM	−0.048	−0.020	−0.020	−0.023	−0.023	−0.018	−0.024	−0.020	−0.013	−0.016	−0.008	−0.010	−0.005	−0.009	−0.010	−0.010
	GGUM	−0.021	−0.002	−0.006	−0.009	−0.018	−0.011	−0.015	−0.019	*−0.025*	*−0.003*	*0.000*	*−0.018*	–	–	–	–
Both skewed predictor and outcome	Baseline	−0.003	−0.009	−0.008	−0.006	−0.009	−0.006	−0.006	−0.005	−0.014	−0.007	−0.006	−0.005	−0.002	−0.004	−0.008	−0.006
	CTT	−0.015	−0.022	−0.021	−0.026	−0.019	−0.022	−0.016	−0.013	−0.039	−0.025	−0.027	−0.025	−0.012	−0.017	−0.018	−0.013
	FA	−0.002	−0.009	−0.008	−0.013	−0.010	−0.015	−0.008	−0.006	−0.027	−0.010	−0.012	−0.011	−0.004	−0.008	−0.010	−0.005
	SGRM	−0.022	−0.023	−0.025	−0.030	−0.023	−0.024	−0.020	−0.017	−0.029	−0.015	−0.016	−0.015	−0.011	−0.012	−0.014	−0.011
	GGUM	−0.005	−0.007	−0.012	−0.013	−0.011	0.003	−0.002	−0.006	*−0.065*	*−0.023*	*−0.029*	*−0.015*	–	–	–	–

Note. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.

TABLE B
R-Squared Change for Different Simulation Conditions

Response data		Ideal point data								Dominance data							
Scale length		15 items				30 items				15 items				30 items			
Sample Size		250	500	1000	2000	250	500	1000	2000	250	500	1000	2000	250	500	1000	2000
No skewness	Baseline	0.021	0.018	0.020	0.019	0.023	0.019	0.019	0.018	0.018	0.018	0.019	0.017	0.023	0.021	0.020	0.019
	CTT	0.006	0.006	0.005	0.004	0.009	0.006	0.005	0.004	0.013	0.005	0.007	0.005	0.018	0.018	0.015	0.015
	FA	0.008	0.006	0.006	0.005	0.010	0.006	0.007	0.006	0.013	0.006	0.007	0.005	0.018	0.018	0.015	0.015
	SGRM	0.007	0.007	0.006	0.005	0.010	0.007	0.007	0.006	0.013	0.006	0.007	0.005	0.018	0.018	0.016	0.016
	GGUM	0.019	0.016	0.017	0.016	0.021	0.018	0.018	0.017	0.017	0.016	0.016	0.015	—	—	—	—
skewed predictor only	Baseline	0.019	0.016	0.015	0.015	0.018	0.017	0.018	0.017	0.017	0.015	0.018	0.017	0.018	0.017	0.016	0.016
	CTT	0.007	0.006	0.005	0.004	0.008	0.006	0.005	0.005	0.010	0.008	0.009	0.008	0.012	0.012	0.010	0.010
	FA	0.007	0.006	0.006	0.005	0.008	0.006	0.006	0.006	0.010	0.008	0.009	0.008	0.012	0.012	0.010	0.010
	SGRM	0.007	0.006	0.005	0.005	0.008	0.006	0.006	0.006	0.011	0.009	0.010	0.009	0.013	0.013	0.011	0.012
	GGUM	0.015	0.014	0.014	0.013	0.016	0.015	0.015	0.015	*0.008*	*0.007*	*0.006*	*0.007*	—	—	—	—
Skewed outcome only	Baseline	0.014	0.016	0.015	0.015	0.015	0.014	0.013	0.013	0.016	0.014	0.016	0.016	0.019	0.016	0.015	0.015
	CTT	0.006	0.006	0.005	0.004	0.008	0.006	0.004	0.004	0.011	0.009	0.011	0.010	0.016	0.013	0.012	0.012
	FA	0.008	0.006	0.006	0.004	0.010	0.008	0.006	0.006	0.011	0.009	0.011	0.010	0.016	0.013	0.012	0.012
	SGRM	0.006	0.006	0.005	0.004	0.010	0.008	0.006	0.006	0.011	0.010	0.011	0.010	0.016	0.014	0.013	0.012
	GGUM	0.012	0.012	0.011	0.011	0.015	0.013	0.012	0.012	*0.008*	*0.009*	*0.009*	*0.008*	—	—	—	—
Both skewed predictor and outcome	Baseline	0.020	0.017	0.017	0.017	0.018	0.017	0.016	0.016	0.016	0.017	0.016	0.016	0.019	0.018	0.016	0.016
	CTT	0.008	0.006	0.005	0.005	0.008	0.006	0.005	0.005	0.008	0.009	0.008	0.008	0.013	0.011	0.009	0.010
	FA	0.009	0.007	0.006	0.005	0.009	0.006	0.006	0.005	0.008	0.009	0.008	0.008	0.013	0.011	0.009	0.010
	SGRM	0.008	0.006	0.005	0.005	0.009	0.006	0.005	0.005	0.009	0.010	0.009	0.009	0.014	0.013	0.011	0.012
	GGUM	0.015	0.013	0.012	0.012	0.015	0.015	0.014	0.013	*0.005*	*0.007*	*0.005*	*0.007*	—	—	—	—

Note. Results were based on the average of 200 iterations, except that the numbers in italics were based on 20 iterations. Baseline = simulated thetas used as predictors; CTT = classical test theory scores; FA = factor analysis scores; SGRM = Samejima's graded response model person scores; GGUM = generalized graded unfolding model person scores.