Check for updates

Article

Organizational Research Methods
2020, Vol. 23(3) 457-486
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1094428119833158
journals.sagepub.com/home/orm
SAGE

Advancing and Evaluating IRT Model Data Fit Indices in Organizational Research

Christopher D. Nye[1], Seang-Hwane Joo[2], Bo Zhang[3], and Stephen Stark[4]

Abstract

Item response theory (IRT) models have a number of advantages for developing and evaluating scales in organizational research. However, these advantages can be obtained only when the IRT model used to estimate the parameters fits the data well. Therefore, examining IRT model fit is important before drawing conclusions from the data. To test model fit, a wide range of indices are available in the IRT literature and have demonstrated utility in past research. Nevertheless, the performance of many of these indices for detecting misfit has not been directly compared in simulations. The current study evaluates a number of these indices to determine their utility for detecting various types of misfit in both dominance and ideal point IRT models. Results indicate that some indices are more effective than others but that none of the indices accurately detected misfit due to multidimensionality in the data. The implications of these results for future organizational research are discussed.

Keywords

item response theory, quantitative research, measurement models, survey research

Item response theory (IRT) models have several advantages for organizational research. In particular, they can be useful for improving the measurement of organizational variables (Carter et al., 2014) and for detecting bias across groups or over time (Tay, Meade, & Cao, 2015). To apply these techniques, a number of IRT models are available for representing the response process to a set of items given the characteristics of the items and the individual's level of the latent trait. However, the advantages of these models are only obtained when the model adequately describes the response process. As a result, a great deal of research has been devoted to examining IRT model-data fit. The vast majority of this research has been conducted using *dominance* models such as the one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL).

[1]Michigan State University, East Lansing, MI, USA
[2]KU Leuven, Leuven, Belgium
[3]University of Illinois at Urbana-Champaign, Champaign, IL, USA
[4]University of South Florida, Tampa, FL, USA

Corresponding Author:
Christopher D. Nye, Department of Psychology, Michigan State University, 316 Physics Rd., East Lansing, MI 48824.
Email: nyechris@msu.edu

Dominance models assume a monotonically increasing relationship between response probabilities and trait levels and are widely used in the organizational literature. However, a different class of models, known as ideal point models, are growing in popularity and have proven useful for modeling response processes for some organizational constructs (e.g., Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009). Therefore, more research is needed to understand IRT model fit for the full range of potential models and constructs examined in organizational research.

To examine model data fit, a number of indices have been proposed and evaluated in the IRT literature. However, most of these indices are variations of the chi-square test and, therefore, will be influenced by the sample size and other characteristics of the data (e.g., frequency distributions of responses). In contrast to the structural equation modeling (SEM) literature, alternatives to the chi-square test are not widely used in IRT research. Although alternative indices are starting to be developed (e.g., Maydeu-Olivares & Joe, 2014), many of these indices have not yet been thoroughly evaluated to determine their accuracy under various conditions or to compare their utility to other alternative indices. Therefore, the present study examined nine different IRT fit indices under various conditions to identify the most useful indicators of model-data misfit.

IRT in Organizational Research

IRT has several methodological advantages over other techniques available to organizational researchers. Although IRT has been used in a number of articles in the organizational literature (see Foster, Min, & Zickar, 2017, for a review), we believe this methodology could be used more frequently to address important issues in this literature. For example, this methodology can be used for scale construction and evaluation, examining measurement bias, and detecting aberrant responding. Table 1 provides a summary of these applications along with some relevant citations and advantages over alternative approaches.

Scale Construction and Evaluation. One of the most common uses of IRT is for scale construction and evaluation. Here, IRT is particularly useful for examining the quality of an existing measure or developing new measures that can assess organizational constructs more effectively. IRT has several unique advantages for scale construction and evaluation over alternative methods. First, IRT can be used to evaluate the quality of a scale regardless of the sample that is used to validate it. In IRT, the quality of items can be evaluated using the item parameters estimated from the model (see below for a more comprehensive discussion of relevant item parameters), which typically include the item difficulty (i.e., how much of the latent trait is needed to endorse an item or get it correct) and discrimination (i.e., how well the item discriminates between high and low levels of the latent trait). A key assumption of IRT models is that these parameters are invariant across different subpopulations of examinees, meaning that they can be readily compared after a linear transformation that puts them on a common scale. In other words, the item difficulty and discrimination will be the same regardless of the sample they are estimated in after applying a linear transformation. This is not the case with traditional methods based in classical test theory (CTT), which involve alternative estimates of difficulty (i.e., the proportion of people endorsing an item) and discrimination (e.g., item-total correlations). Therefore, a set of items in a measure can appear highly discriminating and include a range of difficulties (i.e., indicating a high quality assessment) in one sample but then appear more or less discriminating and difficult in another sample using CTT techniques.

Other advantages of IRT for scale construction and evaluation involve the types of assessments that can be created and constructs that can be assessed effectively. For example, given the assumption of subpopulation invariance described above, IRT can be used to create computer adaptive tests

Table 1. Summary of Applications of IRT in the Organizational Literature.

Methodological Issues	Applications of IRT	IRT Advantages Over Alternative Methods	Relevant Citations
Scale construction and evaluation	• Can be used to shorten a scale or examine the quality of existing scales • Can be used to develop alternative types of assessments.	• In contrast to CTT, IRT item parameters are invariant across samples. • IRT can be used to create adaptive tests but CTT cannot. • Although CTT methods are consistent with dominance models, IRT can be used to create ideal point measures.	• Borman et al. (2001) • Carter, Dalal, Lake, Lin, & Zickar (2011) • Chernyshenko, Stark, Drasgow, & Roberts (2007) • Roznowski (1989) • Tay, Drasgow, Rounds, & Williams (2009)
Identify item and test bias	• Can differentiate observed mean differences from bias • Can detect compensatory DIF and DTF • Can examine differences at the option level	• Comparing mean differences with CTT confounds bias with true differences in the latent trait. • The methods for examining DTF are more widely developed than in CFA.	• Chan, Drasgow, & Sawin (1999) • Nye, Newman, & Joseph (2010) • Raju, Laffitte, & Byrne (2002) • Stark, Chernyshenko, & Drasgow (2004) • Tay, Meade, & Cao (2015)
Detecting aberrant responding	• Can be used to detect different types of aberrant responding such as IER, faking, and spuriously low responding.	• IRT methods can detect multiple types of aberrant responding while other methods can only detect careless responding. • IRT methods are often more effective at detecting aberrant responding than traditional methods of insufficient effort responding (IER).	• Drasgow, Levine, & Zickar (1996) • Stark, Chernyshenko, Chan, Lee, & Drasgow (2001) • Zickar, Gibby, & Robie (2004) • Zickar & Robie (1999)

(CAT), which have been shown to be useful for employment testing (Drasgow et al., 2012; Stark et al., 2014) and for performance appraisals (Borman et al., 2001). With CAT, the items that an individual sees will depend on his or her answers to previous items in the assessment. In other words, the test is adaptive in the sense that test-takers are administered items with characteristics that match their level of the latent trait. Because each individual sees a different set of items, this type of assessment cannot be administered using traditional scale development methods. In addition to their advantages for individual assessment, CATs can also improve the efficiency of the assessment by allowing shorter tests, with some research indicating that adaptive tests can be approximately 50% shorter than static nonadaptive tests while maintaining the same quality of measurement (Stark, Chernyshenko, Drasgow, & White, 2012).

Finally, some types of constructs can be assessed more effectively using IRT than traditional scale development methods. Most measures in the organizational literature have been developed using what is known as a dominance response process. Dominance models assume that the probability of individuals endorsing an item increases monotonically with corresponding increases in their level of the latent trait and the distance between their standing on the trait and the location of the item. Dominance models are consistent with CTT and factor analytic approaches to scale development and are widely used in the organizational literature (Chernyshenko, Stark, Drasgow, & Roberts, 2007). Drasgow, Chernyshenko, and Stark (2010) argued that dominance models are most appropriate for measures that assess an individual's maximal performance and ability in a particular domain. As such, cognitive ability provides a compelling example of a construct that fits within the dominance framework. Individuals are more likely to answer items on a cognitive ability test correctly if their ability is higher than the item's location on the continuum of the latent trait.

Despite the popularity of dominance models, they do not necessarily provide the best description of the response process for all constructs. For measures that do not assess maximal performance and ability, the concept of dominance and the assumptions of this framework may not be appropriate. In these cases, another class of IRT models, known as ideal point models, may provide a better fit to the data. In some areas of research (e.g., political science), the term "ideal point" is used to refer to an individual's standing on the latent trait in the dominance framework (Bafumi, Gelman, Park, & Kaplan, 2005). In the current study, we focus on ideal point models as a separate class of models that are distinct from dominance models and describe a different response process. Ideal point models assume that individuals will tend to endorse an item if the item's location on the latent trait continuum matches their own standing on the latent trait. In other words, if an item is too extreme, or not extreme enough, to match their standing on the trait measured by the assessment, then individuals are less likely to endorse the item.

Drasgow et al. (2010) suggested that ideal point models are best suited for measures that require introspection and self-reflection. The concept behind ideal point models was first proposed by Thurstone (1928) in the context of attitude measurement, which requires individuals to reflect on their thoughts and feelings about a particular topic.[1] Although Thurstone proposed this idea nearly 90 years ago, ideal point models have only recently gained traction in organizational research. Part of the reason for the growing interest in ideal point models is the relatively recent research demonstrating the advantages of ideal point models for a number of constructs. For example, ideal point models have been shown to be useful for understanding employee personality (Carter et al., 2014; Stark et al., 2006), vocational interests (Tay et al., 2009), person-organization fit (Chernyshenko, Stark, & Williams, 2009), performance ratings (Borman et al., 2001), and job attitudes (Carter & Dalal, 2010). In addition, as suggested by Drasgow et al. (2010), these methods will be applicable to any constructs that require employees to think and respond about themselves.

Although the distinction between dominance and ideal point models may seem trivial, past research has suggested that the choice between these models can have important effects on research results and applied decisions. For example, fitting a dominance model to ideal point data can affect the scale development process by influencing the types of items that are retained for a measure (Chernyshenko et al., 2007; Roberts, Laughlin, & Wedell, 1999). Applying an inappropriate model can also result in inaccurate estimates of the latent trait that will affect the rank order of individuals in a sample (Roberts et al., 1999) and correlations between variables (Carter et al., 2014). In addition, applying an incorrect model can also reduce the utility of employee selection decisions (Dalal & Carter, 2015). Therefore, using an inappropriate model for scale construction can have important implications for organizational research and the conclusions that are drawn from it.

Identify Item and Test Bias. Another advantage of IRT for organizational research is the ability to detect item/test bias. In the IRT literature, these methods are known as differential item functioning (DIF) and differential test functioning (DTF). Recent research has provided an extensive review of these techniques (e.g., Tay et al., 2015) and, therefore, we do not discuss the methodology again here. However, we do note several advantages of the IRT approach over alternative methods that are commonly used in the organizational literature.

First, it is important to note that observed mean-level differences do not necessarily reflect test bias and conclusions about these differences will be confounded with any bias in the measure when using CTT techniques. Stark, Chernyshenko, and Drasgow (2004) noted that observed differences between groups are equal to the sum of both bias and impact. Here, impact refers to true differences between groups. For example, if comparing job satisfaction across samples from different countries, any differences that are observed could be due to true differences in satisfaction across countries or due to bias in the measure. Without first testing for bias, the conclusions drawn from these comparisons may be misleading. Moreover, because of these relationships, significant mean differences do not necessarily indicate that there is bias in the measure. Nye and Drasgow (2011) showed that bias in a measure can inflate mean differences or even reverse the direction of these differences in organizational research.

Given the relationship between observed mean differences and bias, IRT provides a more appropriate way of identifying and testing biases in the measure. However, another alternative to IRT is to use confirmatory factor analysis (CFA). The CFA approach is widely used in the organizational literature (see, e.g., Vandenberg & Lance, 2000) and is more commonly known given that it can be estimated using widely available structural equations modeling (SEM) software. Despite this popularity, IRT also has several advantages over CFA tests of bias. For example, one advantage of IRT over the CFA approach is that methods of examining DTF are more widely developed. With DTF, bias at the item level can cancel out when aggregated to the level of the test. This occurs when different items are biased in opposite directions (i.e., one item is biased against one group and another item is biased against the other group). Therefore, DTF examines whether the test as a whole is biased against either group. This application is particularly useful for employee selection testing where researchers may be interested in whether a particular test will result in biased selection decisions if used in this context. Although the concept of DTF is well developed in IRT (see, e.g., Meade, 2010; Stark et al., 2004), similar tests are not available in CFA approaches to detecting bias.

Faking and Insufficient Effort Responding. A substantial amount of recent research has examined the issue of insufficient effort responding (IER; also known as careless responding) in the organizational literature (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Huang, Liu, & Bowling, 2015; Meade & Craig, 2012). This research has identified a number of ways of detecting IER. However, the recommendations for identifying this type of responding suggest methods that may not always be available in organizational research. For example, one recommendation is to use response times to detect careless responding. However, this method will only be available for computerized testing and may not be available at all for archival data. Similarly, another promising approach is to use an index of consistent responding that examines the correlations between items that are theoretically negatively related (Huang et al., 2012). Nevertheless, this approach requires different scales that are negatively correlated to be assessed. There are other options available but the results have been mixed with some studies suggesting that these indices identify individual differences rather than random responding (Goldberg & Kilkowski, 1985) and other studies showing that the various indices generally do not agree on identifying insufficient effort or are insensitive to careless responding (Huang et al., 2012; Meade & Craig, 2012). Finally, these indices are only useful for detecting careless responding and may not be able to detect other types of response biases due to acquiescence, socially desirable responding or faking (i.e., resulting in spuriously high test scores), or spuriously low responding (e.g., due to faking bad, misreading test items, or random response errors such as mistyping an answer).

In contrast, IRT provides another method of detecting IER. A number of studies have been conducted on person-fit indices in the IRT literature. Although a number of indices have been proposed (see Meijer & Sijtsma, 2001, for a review), these indices generally use IRT models to estimate how individuals should be responding given their levels of the latent trait and then compare their predicted response patterns to their observed patterns of responses. Drasgow, Levine, and McLaughlin (1987) conducted a simulation study and found that some IRT person-fit indices could detect aberrant responding nearly 100% of the time, though the power to detect these effects did vary across conditions. Similarly, Meijer and Sijtsma (2001) conducted a qualitative review of IRT person-fit indices and found that detection rates were above 75% for some indices. More recently, Stark et al. (2017) conducted a simulation study that evaluated person-fit indices for specifically detecting IER and faking. Again, these authors found that IRT aberrance indicators could detect both IER and faking nearly 100% of the time in many conditions. In all of these studies, the ability to detect aberrant responding was substantially higher for the IRT indices than in research on traditional methods of detecting IER.

Applying IRT to Organizational Research

Given the potential advantages of IRT, these methods may be particularly useful for advancing organizational research. Therefore, before discussing IRT model fit, we first provide a brief tutorial on fitting IRT models to organizational data. Here we discuss some of the most widely used models in the organizational literature, potential data considerations for applying IRT in empirical research, and the evaluation of model-data fit. The purpose of this tutorial is to help facilitate the use of IRT models in organizational research.

Step 1: Identify an Appropriate IRT Model. The first step to using IRT for organizational research is to identify an appropriate IRT model for the measures used in a study. To do so, researchers must first determine which model is conceptually most appropriate given the data that will be collected and the response format of the measure. As described above, there are two broad classes of IRT models that are widely used in the organizational literature. Probably the most widely used class of models are dominance models. Again, these models assume that the probability of individuals endorsing an item increases monotonically with corresponding increases in their level of the latent trait and the distance between their standing on the trait and the location of the item.

A number of dominance models are available for modeling the response process. For dichotomous responses, one of the most widely used models is the three-parameter logistic model (3PL). This model is described by the equation below:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability of endorsing item i at a given level of the latent trait (θ). In addition, a_i is the discrimination parameter, b_i is the difficulty (or location) parameter, c_i is the guessing parameter, and D is the scaling constant 1.702. This model can be simplified by assuming that there is no guessing (i.e., $c_i = 0$), which results in the two-parameter logistic (2PL) model.

For polytomous items (i.e., items with more than two response options like Likert-type scales), another dominance model is Samejima's (1969) graded response model (GRM). This model indicates the probability of endorsing a particular response option using the following equation:

$$P_{ik}(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik-1})}} - \frac{1}{1 + e^{-a_i(\theta - b_{ik})}}$$

where $P_{ik}(\theta)$ is the probability that an individual with a given level of the latent trait (θ) will endorse option k on item i, and b_{ik} is the location (or difficulty) parameter for the corresponding option.[2] For this model, there will be C – 1 b_{ik} parameters, where C is the total number of response options.

In contrast to dominance models, ideal point models assume that individuals will tend to endorse an item if the item's location on the latent trait continuum matches their own standing on the latent trait. In the organizational literature, the most widely used ideal point model is the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), which has been used in several organizational studies (e.g., Carter et al., 2014; Stark et al., 2006; Tay et al., 2009) and is potentially applicable to a number of other areas in organizational research (Drasgow et al., 2010). For dichotomous items (readers are referred to Roberts et al., 2000, for a description of GGUM for polytomous items), this model is defined as

$$P(X_i = 1|\theta) = \frac{\exp\Big(a_i[(\theta-\delta_i)-\tau_{i1}]\Big)+\exp\Big(a_i[2(\theta-\delta_i)-\tau_{i1}]\Big)}{1+\exp\Big(a_i[3(\theta-\delta_i)]\Big)+\exp\Big(a_i[(\theta-\delta_i)-\tau_{i1}]\Big)+\exp\Big(a_i[2(\theta-\delta_i)-\tau_{i1}]\Big)}$$

where a_i is the discrimination parameter for item i, δ_i is the location of item i on the latent trait continuum, and τ_{i1} is the location of the response threshold on the latent trait continuum.

Step 2: Collecting Data and Fitting IRT Models. Once an appropriate IRT model has been identified, the next step is to collect the data and/or estimate the IRT parameters based on the model being evaluated. In this stage of the process, there are a number of factors that need to be considered to obtain accurate estimates of IRT parameters. One issue that needs to be considered is the sample size. IRT estimation generally requires larger sample sizes than other approaches. However, there is no one answer for how large the sample size needs to be. Instead, the sample size requirements will depend on a number of factors, including the model that is being estimated. Nevertheless, some have suggested that at least 500 individuals are needed to get accurate estimates of the item parameters for the 2PL and GRM models (Embretson & Reise, 2000; Stark et al., 2006). These sample size requirements are often quite a bit larger than recommended sample sizes for other approaches such as CFA.

In addition to larger sample sizes, IRT models also generally make several assumptions about the measures that are being analyzed. One assumption is that the data have the appropriate dimensionality for the IRT model. All of the models described above assume the response data are unidimensional. However, multidimensional generalizations are available for most of these models (Forero & Maydeu-Olivares, 2009; Reckase, 2009). Therefore, it is important to first examine the dimensionality of a measure prior to estimating IRT parameters. This can be done using standard software packages for estimating a CFA model. However, research has also indicated that the data do no need to be strictly unidimensional (see Foster et al., 2017, for a review of this research). Instead, the data just need to be "unidimensional enough." For example, Reckase (1979) showed that IRT estimates were generally accurate when the first factor accounted for at least 20% of the variance. Similarly, Drasgow and Lissak (1983) found that IRT estimates are still accurate as long as the latent factors are highly correlated (e.g., above .50).

All of the IRT models described above also assume local independence. Local independence means that all of the items being assessed are uncorrelated after accounting for the latent trait they assess. In other words, the only reason that the items are correlated is because of their relationship with the latent trait. In practice, this assumption suggests that there are no correlated errors among the items. Consequently, this assumption can also be tested using common SEM software to estimate a CFA model.[3]

After ensuring a sufficient sample size for analyses and verifying that the data meet the assumptions of the IRT model, item parameters can be estimated using a number of different software packages. In the past, estimating IRT models required specialized software packages that were primarily used for IRT modeling and did not estimate other more commonly used statistical models. This was a major disadvantage of using IRT for organizational research because few people were familiar with these software packages or their use. However, this is no longer the case and IRT models can now be estimated in several widely used statistical packages, including Mplus (Muthén & Muthén, 1998-2017) and R, which are also used for other statistical analyses. Therefore, researchers who are familiar with this software can easily apply IRT models using widely used statistical packages. Example R code for estimating the IRT models examined in the present study is provided in the online supplemental material and at the following website: https://psychology.psy.msu.edu/pers_nye/IRT_syntax/.

Step 3: Evaluate IRT Model Fit. After identifying an appropriate IRT model and estimating IRT parameters, the next step is to evaluate the fit of the model to the data. Although evaluating model fit is ubiquitous in the SEM literature, this step is frequently ignored when applying IRT to organizational research. For example, in their review of the literature, Foster et al. (2017) noted that over 40% of articles using IRT in the organizational literature have not reported any information about fit. However, given the number of IRT models available, identifying the correct model for characterizing the response process in organizational research is particularly important (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). In fact, past research has demonstrated that selecting an inappropriate IRT model can have serious consequences for scale construction (Chernyshenko et al., 2007; Roberts et al., 1999), test score equating (e.g., Bolt, 1999; Kaskowitz & De Ayala, 2001), parameter estimation (e.g., DeMars, 2005), adaptive testing (e.g., De Ayala, Dodd, & Koch, 1992), and differential item functioning (e.g., Bolt, 2002). As noted above, applying an inappropriate IRT model can also result in inaccurate estimates of the latent trait that will affect the rank order of individuals in a sample (Roberts et al., 1999) and correlations with external variables (i.e., validity; Carter et al., 2014). Therefore, testing for IRT model fit is critical because the benefits of IRT described above cannot be realized if the model does not fit the data well.

A number of fit indices have been developed for testing IRT model fit. One way to operationalize model fit is to examine the extent to which the observed proportions of correct answers on a measure match the model predicted proportions using a chi-square fit statistic. One such index, known as Yen's Q_1 (*Y-Q_1*; Yen, 1981), is defined as

$$Y\!-\!Q_{1i} = \sum_{r=1}^{10} \frac{N_r(O_{ir} - E_{ir})^2}{E_{ir}(1 - E_{ir})}.$$

Building on Bock's (1972) chi-square index, Y-Q_1 is calculated by rank ordering examinees using their trait estimates (i.e., $\hat{\theta}$) and then dividing them into 10 score groups of approximately equal size. N_r is the number of examinees in cell r, O_{ir} is the observed proportion of examinees in cell r that correctly answers item i, and E_{ir} is the proportion of examinees in cell r that are predicted to answer item i correctly. As such, Y-Q_1 is essentially the sum of standardized residuals and follows a χ^2 distribution with $10 - g$ degrees of freedom, where g is the number of item parameters estimated.

Although Yen (1984) found that Y-Q_1 had some difficulty identifying misfit when the 2PL model was fit to 3PL data, her study also indicated that Y-Q_1 is sensitive to other forms of model misfit. However, in the context of the present study, Y-Q_1 has two notable limitations. First, some have questioned the appropriateness of dividing scores into an arbitrary number of groups based on trait estimates. The problem is that this approach can provide inaccurate estimates of the differences between the observed and expected proportions. The reason for these inaccuracies is that allocating individuals to score groups of equal size will depend on the sample being used for estimation because the cut points and number of intervals that are used will vary depending on the distribution of the latent trait. This can result in high Type I error rates under some conditions (Orlando & Thissen, 2000). A second limitation of this index is that it can result in inaccurate estimates of fit for some types of models. As Roberts et al. (2000) pointed out, the ideal point response process can result in small expected frequencies for some response categories. Although small expected frequencies can be a problem for any IRT model, this issue is particularly salient for ideal point models because these models assume that individuals tend to endorse items that are close to their level of the latent trait. As a result, the pattern of these small expected frequencies makes it difficult to correct this issue by combining response categories, which is a common way to address this problem for other IRT models. Given these limitations, more research is needed to evaluate *Y-Q₁* as an index of IRT model fit.

Drasgow, Levine, Tsien, Williams, and Mead (1995) suggested another family of χ^2 statistics that does not require examinees to be allocated to separate score groups. Instead, observed and expected frequencies are summed over the number of response options. Using this formulation, the χ^2 for item singles and doubles can be calculated. For m items, m and $\binom{m}{2}$ χ^2 statistics can be computed for item singles and doubles, respectively. As described by Drasgow et al., the expectations for response option k, where $k = 0{,}1$, on item i and item double, i, j, are given by

$$E_i(k) = N\int P(u_i = k|\theta)f(\theta)d\theta$$

and

$$E_{i,j}(k,k') = N\int P(u_i = k|\theta)P(u_j = k'|\theta)f(\theta)d\theta,$$

respectively. Here, f is the ability density, which is usually taken as normal (0,1), and u denotes a scored response. The χ^2 is then computed with the usual formula

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

and the χ^2 for doubles is computed using the formula for a 2 x 2 contingency table. This family of chi-squares can be extended to triples of items and beyond. However, the expected and observed frequencies of many cells quickly diminish as the size of the contingency table increases and the power of the chi-square would be expected to decline as a result. The chi-squares for item singles, doubles, and triples are typically evaluated by examining the ratio of the chi-square to the degrees of freedom (*df*; Drasgow et al., 1995). As a useful heuristic, values of the χ^2/df ratio less than 3 are considered indicative of good model-data fit (Chernyshenko et al., 2001).

The chi-squares for singles, doubles, and triples have been used to assess the fit of both dominance and ideal point models in previous research (Chernyshenko et al., 2001; Drasgow et al., 1995; Stark et al., 2006) and some models have been found to fit and others misfit. In addition, two recent simulation studies have examined the accuracy of these chi-square indices. In one study, Tay, Ali, Drasgow, and Williams (2011) examined the χ^2/df ratios for item doubles and triples and compared the efficacy of these indices for detecting misfit. Their results indicated that the χ^2/df ratios were generally effective at identifying the correct IRT model but, in some cases, had trouble identifying misfit when the GGUM was fit to data generated from the 2PL model. In another study, Tay and Drasgow (2012) examined the χ^2/df ratios for detecting misfit in dominance models across a range of conditions and found that this index had difficulty distinguishing between some dominance models (e.g., 2PL and 3PL).

One limitation of chi-square fit indices in general is that they are severely influenced by sample size. To address this issue for the chi-square indices proposed by Drasgow et al. (1995), some have suggested examining chi-square values that have been adjusted to a smaller sample size when N is large (e.g., greater than 3,000; Chernyshenko et al., 2001; Tay et al., 2011). Although these adjustments can help to address problems with the increasing chi-square values that result from large sample sizes, some research has indicated that these adjusted χ^2/df ratios may still result in high Type I error rates for smaller sample sizes (Tay & Drasgow, 2012).

Problems with chi-square fit indices have been encountered in the structural equation modeling (SEM) literature as well. To address these issues in SEM, alternative fit indices have been proposed and are commonly used to evaluate model-data fit. However, until recently, similar fit indices were not available for IRT models. Recent work has focused on addressing this issue and developing new IRT fit indices that can supplement the chi-square fit statistics and address their potential limitations. For example, Maydeu-Olivares and Joe (2014) focused on developing indices of approximate fit and developed a standardized root mean square residual (SRMSR).[4] Based on a set of simulations, these authors suggested that an SRMSR $\leq$ .05 indicates that the model is an acceptable approximation to the data.

Another limitation of chi-square fit indices is that the number of potential response patterns increases substantially as the number of items in a measure increases. For dichotomous items, there will be 2^{m} potential response patterns, where m is the number of items in the measure. As a result, it is nearly impossible to calculate a chi-square to compare observed to empirical probabilities for a full contingency table with 2^{m} cells. To address this issue, most chi-square fit indices examine lower-order approximations (e.g., for single items, doubles, or triples) of the full contingency table. However, another alternative is to use a Bayesian approach to model-checking that does not require examining such lower-order approximations. For example, an approach called the posterior predictive model check (PPMC; Rubin, 1984) has shown promise in recent research (Li, Bolt, & Fu, 2006; Sinharay, Johnson, & Stern, 2006; Zhu & Stone, 2012). This approach examines model-data fit by comparing observed data to replicated data that are generated from the model being evaluated. The replicated data are then used to evaluate the probability of the observed data by calculating posterior predictive p- (PPP) values to determine the extremity of the observed data given the posterior distribution. In other words, the question is whether the observed data look like the replicated data generated from the model. The fit of multiple models can be evaluated by comparing the number of items (or item doubles) with PPP values near .50, which indicates that there is no systematic difference between the observed and replicated data (Sinharay et al., 2006). A model that results in more items with PPP values near .50 compared to other models would be considered a good fitting model.

Another approach to examining model data fit is to examine Akaike's information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). Both the AIC and BIC have been widely used as model fit indices and provide information about model fit relative to the number of parameters in the model. Although the AIC and BIC have shown some promise for correctly identifying the fit of the 1PL and 2PL IRT models (Kang, Cohen, & Sung, 2009), these indices have some limitations for some types of IRT models. For example, these indices should not be used with MCMC estimation because this method uses a prior distribution which creates a dependence between the parameters that makes estimating the number of parameters difficult. Because the calculation of AIC and BIC depend on the number of parameters in the model, these indices may be inaccurate with MCMC estimation (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). This is particularly important given that recent research has demonstrated that MCMC estimation can provide more accurate parameter estimates for some ideal point models (de la Torre, Stark, & Chernyshenko, 2006; Wang, 2014).

The Present Study

Given the importance of IRT model-data fit and the various indices that are available, examining these indices and comparing their performance under a broad range of conditions is important for understanding their utility in organizational research using IRT methods. Although each of these indices has shown promise in previous research, we are not aware of any studies that have compared the accuracy of these indices to each other or examined their utility for identifying misfit in a broad range of IRT models. In other words, few studies have examined all of these indices simultaneously and compared their performance under the same conditions. In addition, even fewer studies have examined their performance for both dominance and ideal point models. Although previous research has examined model data fit for ideal point models (DeMars, 2004; Tay et al., 2011), these studies have generally focused only on the chi-square fit indices and have not evaluated the recently developed alternatives to these indices. Therefore, the present study examined a broader range of indices using computer simulations to evaluate their performance.

Methods

We conducted a simulation study manipulating four independent variables: the number of items (10, 20, 40), sample size (250, 500, 1,000, 2,000), type of data (dichotomous, polytomous), and model used to generate the data. For dichotomous (i.e., categorical items with two response options) conditions, data were generated from the 2PL model, 3PL model, or GGUM. For polytomous (i.e., Likert-type items with four response options) conditions, data were generated from the graded response model (GRM) or the GGUM. In addition to these unidimensional models, we also simulated multidimensional data in both the dichotomous (i.e., multidimensional 2PL [M2PL]) and polytomous (i.e., multidimensional GRM [MGRM]) conditions to examine misfit due to violations of the IRT assumption of unidimensionality. Table 2 provides a description of each of the models simulated in this study along with a summary of their applications. Using these models, the present study evaluated the performance of the fit indices examined here across 84 simulated conditions with 100 replications in each condition for a total of 8,400 samples. All simulations and analyses were conducted in Ox (Doornik, 2009), including the data generation, model fitting, and calculation of fit indices.[5]

Data Generation

Generating Parameters. The item parameters used to generate the data from specific IRT models were randomly generated from distributions of parameters obtained from previous simulation and substantive IRT studies (e.g., Roberts et al., 2000; Stark et al., 2006). For dichotomous conditions, the discrimination parameters used to generate the data for the 2PL, 3PL, GGUM, and M2PL models were obtained from a random uniform distribution [0.5, 2]. The location parameters used to generate the data for these models were obtained from a random uniform distribution [–2, 2]. Also, the generating threshold parameter (τ) for the GGUM and the guessing parameter, c, for the 3PL model were obtained from random uniform distributions [–1.4, 0.4] and [0, 0.3], respectively. For polytomous conditions, data were simulated with four response categories and the distribution for generating the discrimination parameter was the same as in the dichotomous conditions. In addition, the threshold parameters (b_1, b_2, b_3) for generating data from the GRM and MGRM models were obtained from random uniform distributions [–2, –0.5], [–0.5, 0.5], and [0.5, 2], respectively. The ranges of these uniform distributions were chosen based on past simulation work with the GRM (e.g., Kieftenbeld & Natesan, 2012). For the GGUM, the highest threshold parameter (τ_3) was randomly sampled from a uniform distribution [–1.5, –0.5]; then the thresholds for τ_2 and τ_1 were

Table 2. Summary of the IRT Models That Were Simulated in This Study.

IRT Model	Description	Application
Two-parameter logistic model (2PL)	Follows a dominance response process and assumes that all items vary on discrimination (a_i) and difficulty (b_i).	Used for dichotomous (e.g., yes/no) items where guessing is unlikely (e.g., open-ended questions with many possible answers). Assumes unidimensionality.
Three-parameter logistic model (3PL)	Follows a dominance response process and assumes that all items vary on discrimination (a_i), difficulty (b_i), and guessing (c_i).	Used for dichotomous items where guessing is likely to occur (e.g., items with a right or wrong answer or yes/no response options). Assumes unidimensionality.
Multidimensional two-parameter logistic model (M2PL)	Follows a dominance response process and assumes that all items vary on a vector of discrimination parameters ($a_i = a_1, \ldots, a_m$) and difficulty (b_i).	Used for dichotomous items where guessing is unlikely (e.g., open-ended questions with many possible answers). Can be used when the data are multidimensional.
Graded response model (GRM)	Follows a dominance response process and assumes that all items vary on discrimination (a_i) and each response option varies on difficulty (b_{ik}). In other words, each item will have one a parameter and C – 1 b parameters, where C is the total number of response options.	Used for polytomous (i.e., Likert-type) items with three or more response options. Assumes unidimensionality.
Multidimensional graded response model (MGRM)	Follows a dominance response process and assumes that all items vary on a vector of discrimination parameters ($a_i = a_1, \ldots, a_m$) and each response option varies on difficulty (b_{ik}). Each item will have a vector of a parameters for each latent factor and C – 1 b parameters, where C is the total number of response options.	Used for polytomous (i.e., Likert-type) items with three or more response options. Can be used when the data are multidimensional.
Generalized graded unfolding model (GGUM)	Follows an ideal point response process and assumes all items vary on discrimination (a_i), item location (δ_i), and thresholds (τ_{il}).	Can be used for dichotomous or polytomous ideal point items. Assumes unidimensionality.

generated using a recursive formula, wherein a constant of –.25 and a random error sampled from a $N(0, 0.04)$ distribution were added to τ_3 and τ_2, respectively. This recursive method mirrors the threshold parameters that have been used in previous GGUM simulation studies (e.g., Roberts et al., 2000). The person parameters (θ) used to generate the data were randomly selected from a $N(0, 1)$ distribution for the unidimensional models. For the multidimensional models, the person parameters were generated from a multivariate normal distribution with 0 mean vector, variances of 1, and a covariance of 0.5.

Response Data Generation. Using the item parameters described above, response data were generated by comparing randomly sampled numbers from a uniform (0, 1) distribution to the response probabilities for the respective items. For dichotomous conditions, response data were scored 1 if the response probability was greater than the random uniform number and 0 otherwise. For the polytomous conditions, the response data were scored 3 if a randomly sampled uniform number was greater than the sum of the probabilities for categories 0, 1, and 2. If the random number was less than the sum of the probabilities for categories 0, 1, and 2 but greater than the sum for categories 0 and 1, then the response was scored 2, and so forth.

Model Fitting and Estimation

The simulated data described above were then fit with either a dominance model (2PL for dichotomous conditions and GRM for polytomous conditions) or an ideal-point model (GGUM for both dichotomous and polytomous conditions) to evaluate misfit. The goal of this approach was to examine both situations in which dominance data were fit with an ideal point model and situations in which the opposite was true to determine the sensitivity of the fit indices under both conditions. The 2PL (or GRM for polytomous data) and GGUM models were estimated using Markov chain Monte Carlo (MCMC) estimation with Metropolis-Hasting within Gibbs sampling (Patz & Junker, 1999). The MCMC estimation was implemented in Ox (Doornik, 2009). Prior to estimating the parameters, initial values were specified for each parameter of the models. The initial discrimination parameters for both 2PL (and GRM) and GGUM models were fixed at 1 across all items. The initial location parameters, however, were specified separately for the two fitted models. For the 2PL (and GRM) model, the initial location parameters were fixed at 0 across all items. For the GGUM, the initial location parameters were generated using the method described by Roberts and Laughlin (1996). For polytomous conditions, the initial threshold parameters of the GRM were set at –1, 0, and 1, respectively. The initial values for the GGUM threshold parameters were set at –1 in the dichotomous conditions, and –2, –1, and 0 in the polytomous conditions. For person parameters, the initial values were generated randomly from a $N(0, 1)$ distribution.

For the prior distributions of the MCMC algorithm, a four-parameter beta distribution, *Beta* (ν, ω, a, b), was chosen. In general, the lognormal and normal distributions are used for the item discrimination and difficulty (location or threshold for the GGUM) parameters, respectively. However, the four-parameter beta distribution is flexible and can mimic a wide range of distributions including normal or lognormal by manipulating its shape parameters (ν, ω) and range parameters (a, b). For example, the beta distribution with parameters (12.66, –12.66, –5, 5) can approximate the standard normal distribution with precision (de la Torre et al., 2006). Parameters for the four-parameter beta priors were chosen based on previous MCMC studies (e.g., de la Torre et al., 2006; Joo, Lee, & Stark, 2017; Kim & Bolt, 2007). For the 2PL and GRM, the prior parameters for alpha and beta were (1.5, 1.5, .25, 4) and (2, 2, –5, 5), respectively. For the GGUM, the prior parameters for alpha, delta and tau were (1.5, 1.5, .25, 4), (2, 2, –5, 5), and (2, 2, –6, 6), respectively.

Item and person parameters from the fitted models were estimated by computing the average of the posterior samples after a "burn-in" period. Based on previous research (de la Torre et al., 2006; Joo et al., 2017) and preliminary experiments with the fitted models, we chose the numbers of chains and iterations for the MCMC algorithm to satisfy convergence criteria. 30,000 samples were drawn from three independent chains and the first 10,000 samples in each chain were discarded as burn-in. For a convergence check, the $\hat{R}$ statistic (Gelman & Rubin, 1992) was computed for each item parameter and convergence of the MCMC algorithm was verified based on criteria of $\hat{R} < 1.2$.

Model-Data Fit Indices

For each calibration model, we computed several model-data fit indices: the PPMC, $Y\text{-}Q_1$, Drasgow et al.'s (1995) chi-squares for item singles, doubles, and triples, the standardized root mean square residuals (SRSMR), AIC, and BIC. Here, we examined both the original chi-square/df ratios for item singles, doubles, and triples as well as the sample-size-adjusted chi-square/df values (i.e., adjusted to a sample size of 3,000). All model-data fit indices were computed in each replication using the estimated item and person parameters. For item-level fit statistics (e.g., $Y\text{-}Q_1$, chi-squares for singles, doubles, and triples), overall model-data fit was computed by averaging over the number of items to compare with scale-level fit statistics (e.g., AIC, BIC, PPMC, SRMSR).

We used conventional cutoffs to evaluate fit with each of these indices. For example, the average Yen's Q_1 was compared to a chi-square distribution with $10 - g$ (where g is the number of item parameters estimated) degrees of freedom and significant values indicated model-data misfit. In contrast, the average chi-squares for item singles, doubles, and triples (both the original values and the sample-size-adjusted values) were divided by their degrees of freedom and ratios greater than 3 indicated misfit (Chernyshenko et al., 2001). For the SRMSR, we used values greater than .05 as the criterion for identifying misfit as suggested by Maydeu-Olivares and Joe (2014). For AIC and BIC, we evaluated fit by comparing the values of these indices across the models fit to a particular dataset and selecting the model with the lowest information criterion as the best fitting model. For example, in a condition with dichotomous data, we first fit the 2PL and GGUM models to the 2PL, 3PL, GGUM and M2PL data, respectively. We then calculated AIC and BIC for each of the models and selected the model with the lowest values as the best fitting model for each sample. Across conditions, we report the number of times that the fitted model resulted in the lowest values for the AIC and BIC.

Finally, we also identified the best fitting model using the PPMC (Rubin, 1984) by computing posterior predictive p- (PPP) values. These PPP values were computed by adapting the method described by Sinharay (2006) in which the model is evaluated using the posterior predictive distribution. Due to the complexity of the posterior distribution and the integration procedure, the posterior predictive distributions were obtained using the following process:

1. After the burn-in period, Markov chain samples of item and person parameters were drawn from the posterior distribution.
2. Next, we generated replicated response data, $\boldsymbol{y}^{rep}$, using the item and person parameters sampled from the posterior distributions.
3. Then we computed a discrepancy statistic D_i in both the observed data $\boldsymbol{y}$ and the replicated data $\boldsymbol{y}^{rep}$ for each sample. That is, $D_i = \sum_k \frac{[Obs_{ik} - Exp_{ik}]^2}{Exp_{ik}}$, where Obs_{ik} and Exp_{ik} are the observed and expected number of examinees scoring in response category k on item i, respectively, and Exp_{ik} was computed by summing the probabilities of scoring in response category k across all N examinees: $Exp_{ik} = \sum_{r=1}^{N} P_{rjk}$.
4. Finally, we compared the discrepancy statistic for the replicated data to that of the observed data. The PPP values were then computed as the proportion of the discrepancy statistics for the replicated data that were greater than the corresponding statistics for the observed data across the posterior samples. Extreme PPP values (i.e., less than .05 or greater than .95) indicated model-data misfit (Sinharay, 2006).

Results

To provide information about the distributions of fit indices in these simulations, the means of each fit index are presented in the online supplemental material for each of the conditions simulated here. In addition, the Type I error rates and power for these indices are shown in Tables 3 and 4 for the dichotomous conditions that were fit with either the 2PL model or GGUM, respectively. In these tables, the Type I error rates were determined by calculating the percentage of replications in each condition where a fit index incorrectly identified misfit when the correct model was fit to the data. In contrast, power was determined by calculating the percentage of replications in each condition

Table 3. Power and Type I Error Rates of the Model Fit Indices When Fit With the 2PL Model in the Dichotomous Conditions.

		Data Gen.	Fitted Model (2PL)								
I	*N*	Model	Singles	Doubles	Triples	Singles*	Doubles*	Triples*	Yen's Q1	PPP	SRMSR
10	250	2PL	.00	.00	.00	.00	.00	.00	.05	.01	.00
		3PL	.00	.00	.00	.00	.00	.00	.11	.07	.00
		GGUM	.00	.00	.00	.00	.97	1.00	.49	.25	.96
		M2PL	.00	.00	.00	.00	.00	.00	.12	.18	.00
	500	2PL	.00	.00	.00	.00	.00	.00	.22	.02	.01
		3PL	.00	.00	.00	.00	.00	.00	.23	.06	.01
		GGUM	.00	.00	.06	.00	1.00	1.00	.77	.34	.99
		M2PL	.00	.00	.00	.00	.00	.00	.32	.13	.00
	1,000	2PL	.00	.00	.00	.00	.00	.00	.44	.05	.00
		3PL	.00	.00	.00	.00	.00	.00	.70	.08	.00
		GGUM	.00	.22	1.00	.00	1.00	1.00	.86	.58	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.59	.12	.00
	2,000	2PL	.00	.00	.00	.00	.00	.00	.91	.07	.00
		3PL	.00	.00	.00	.00	.00	.00	.91	.12	.00
		GGUM	.04	1.00	1.00	.04	1.00	1.00	.90	.85	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.85	.07	.00
20	250	2PL	.00	.00	.00	.00	.00	.00	.05	.01	.00
		3PL	.00	.00	.00	.00	.00	.00	.12	.06	.00
		GGUM	.00	.00	.00	.08	1.00	1.00	.47	.35	1.00
		M2PL	.00	.08	.00	.00	.00	.00	.03	.08	.00
	500	2PL	.00	.00	.00	.00	.00	.00	.20	.02	.00
		3PL	.00	.00	.00	.00	.00	.00	.30	.05	.00
		GGUM	.01	.35	.90	.38	1.00	1.00	.74	.38	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.23	.05	.09
	1,000	2PL	.00	.00	.00	.00	.00	.00	.47	.06	.00
		3PL	.00	.00	.00	.00	.00	.00	.64	.08	.00
		GGUM	.07	1.00	1.00	.61	1.00	1.00	.92	.68	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.43	.04	.00
	2,000	2PL	.00	.00	.00	.00	.00	.00	.81	.04	.00
		3PL	.00	.00	.00	.00	.00	.00	.89	.25	.00
		GGUM	.12	1.00	1.00	.64	1.00	1.00	.98	.86	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.69	.06	.00
40	250	2PL	.00	.00	.00	.00	.00	.00	.04	.02	.00
		3PL	.00	.00	.00	.00	.00	.00	.10	.07	.00
		GGUM	.00	.00	.07	.92	1.00	1.00	.45	.31	1.00
		M2PL	.00	.00	.00	.00	.05	.08	.03	.05	.00
	500	2PL	.00	.00	.00	.00	.00	.00	.06	.05	.00
		3PL	.00	.00	.00	.00	.00	.00	.09	.06	.00
		GGUM	.12	.71	.98	1.00	1.00	1.00	.69	.41	.99
		M2PL	.00	.00	.00	.03	.13	.05	.06	.07	.00
	1,000	2PL	.00	.00	.00	.00	.00	.00	.10	.06	.00
		3PL	.00	.00	.00	.00	.00	.00	.22	.09	.00
		GGUM	.17	1.00	1.00	1.00	1.00	1.00	.87	.66	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.08	.06	.00
	2,000	2PL	.00	.00	.00	.00	.00	.00	.23	.04	.00
		3PL	.00	.00	.00	.00	.00	.00	.47	.10	.00
		GGUM	.20	1.00	1.00	1.00	1.00	1.00	.97	.72	1.00
		M2PL	.00	.00	.00	.00	.00	.00	.17	.08	.00

Note: *I* = number of items; *N* = sample size; Data Gen. Models = data generation models. Shaded cells indicate correctly specified models. Singles, doubles, and triples are Drasgow et al.'s (1995) chi-square model-data fit statistics.
*Indicates the sample-size-adjusted model fit index (Chernyshenko, Stark, Drasgow, & Roberts, 2007).

Table 4. Power and Type I Error Rates of the Model Fit Indices When Fit With the GGUM in the Dichotomous Conditions.

		Data Gen.	Fitted Model (GGUM)								
I	*N*	Model	Singles	Doubles	Triples	Singles*	Doubles*	Triples*	Yen's Q1	PPP	SRMSR
10	250	2PL	.43	.39	.38	.53	.52	.51	.73	.54	.61
		3PL	.40	.34	.22	.72	.68	.68	.92	.81	.90
		GGUM	.00	.00	.00	.00	.00	.00	.16	.02	.00
		M2PL	.05	.66	.82	.51	.98	.98	.65	.25	1.00
	500	2PL	.56	.56	.44	.83	.80	.73	.96	.89	.87
		3PL	.57	.41	.33	.75	.67	.66	.96	.78	.89
		GGUM	.00	.00	.00	.00	.00	.00	.21	.05	.00
		M2PL	.12	1.00	1.00	.36	1.00	1.00	.93	.53	1.00
	1,000	2PL	.71	.62	.55	.93	.87	.78	1.00	1.00	.99
		3PL	.79	.70	.61	.87	.84	.83	.99	.89	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.51	.04	.00
		M2PL	.11	1.00	1.00	.47	1.00	1.00	1.00	.83	1.00
	2,000	2PL	.69	.61	.52	.82	.70	.68	1.00	1.00	1.00
		3PL	.76	.68	.69	.81	.76	.78	1.00	.97	.98
		GGUM	.00	.00	.00	.00	.00	.00	.92	.06	.00
		M2PL	.24	1.00	1.00	.82	1.00	1.00	1.00	.98	1.00
20	250	2PL	.59	.46	.34	.93	.85	.77	.95	.93	.91
		3PL	.12	.04	.02	.72	.60	.63	.96	.94	.96
		GGUM	.00	.00	.00	.00	.00	.00	.17	.04	.00
		M2PL	.14	.23	.21	.91	.84	.71	.61	.50	.93
	500	2PL	.73	.58	.44	.97	.92	.88	1.00	.99	.99
		3PL	.15	.04	.03	.68	.54	.54	1.00	.98	.98
		GGUM	.00	.00	.00	.00	.00	.00	.32	.05	.00
		M2PL	.58	.94	.93	1.00	1.00	1.00	1.00	.91	1.00
	1,000	2PL	.90	.69	.48	1.00	.97	.94	1.00	1.00	1.00
		3PL	.26	.16	.09	.80	.54	.58	1.00	1.00	.98
		GGUM	.00	.00	.00	.00	.00	.00	.54	.06	.00
		M2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2,000	2PL	.99	.96	.64	1.00	.98	.96	1.00	1.00	1.00
		3PL	.61	.48	.41	.91	.60	.65	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.82	.06	.00
		M2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	250	2PL	.84	.62	.41	1.00	1.00	1.00	.98	.98	.94
		3PL	.18	.06	.02	.93	.87	.86	.96	.94	.98
		GGUM	.00	.00	.00	.00	.00	.00	.08	.09	.00
		M2PL	.83	.84	.73	1.00	1.00	1.00	.97	.94	1.00
	500	2PL	.94	.62	.42	1.00	1.00	1.00	1.00	1.00	1.00
		3PL	.38	.11	.09	1.00	.98	.94	1.00	.99	.97
		GGUM	.00	.00	.00	.00	.00	.00	.08	.07	.00
		M2PL	.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	1,000	2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		3PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.15	.05	.00
		M2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2,000	2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		3PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.32	.05	.00
		M2PL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: *I* = number of items; *N* = sample size; Data Gen. Models = data generation models. Shaded cells indicate correctly specified models. Singles, doubles, and triples are Drasgow et al.'s (1995) chi-square model-data fit statistics.

where a fit index correctly identified model-data misfit. As shown in Table 3, the performance of the fit indices when evaluating the fit of the 2PL model varied widely and depended on the generating model. Both Yen's Q_1 and the PPP performed poorly when the 2PL model was fit to the data. Although Yen's Q_1 did have moderately high power for detecting misfit when the 2PL model was fit to GGUM data, this index also had Type I error rates as high as .91 under some conditions. In addition, Yen's Q_1 was affected by sample size such that it showed higher power for detecting misfit in larger samples. In contrast, the PPP appeared to be insensitive to all forms of misfit when the 2PL model was fit to the data, except when the data were generated from the GGUM (although Type I error was well-controlled). However, even in these conditions, the power of the PPP was still low. In general, both Yen's Q_1 and PPP performed poorly under these conditions.

Table 3 also shows that the adjusted (for sample size) chi-squares for doubles and triples as well as the SRMSR were good at detecting misfit in some conditions but not in others. For example, each of these indices demonstrated high power for identifying ideal point data that were incorrectly fit with the 2PL model. In these conditions, the adjusted chi-squares for doubles and triples as well as the SRMSR had power near 1.00. As expected, the unadjusted chi-square was affected by sample size such that misfit was more likely to be identified when the sample size was larger. The effect of sample size on the unadjusted chi-square was most evident in the smallest samples where the power dropped as low as .00 under some conditions. However, this chi-square statistic also appeared to be affected by the number of items in the measure because the power of the unadjusted chi-square for doubles was .71 with a sample size of 500 and 40 items compared to just .35 for the same sample size and 20 items. In addition, the unadjusted chi-square for single items was a relatively poor indicator of misfit under most conditions. In contrast, when the chi-square for single items was adjusted for sample size, power was substantially better, though this index was still affected by the number of items and power was still low in many conditions. However, the SRMSR did not appear to be affected by sample size or the number of items and the power to detect misfit for this index was nearly 1.00 under all conditions when the 2PL model was fit to GGUM data.

Despite the relatively high power for detecting misfit when ideal point data were fit with dominance models, neither the chi-squares for item doubles and triples nor the SRMSR accurately detected misfit due to multidimensionality. Similarly, these indices did not accurately detect misfit when the 2PL model was fit to 3PL data. For most of the indices in Table 3, the power under these conditions was at or near zero. In other words, these indices appeared to be insensitive to conditions in which too few parameters were estimated or the assumption of unidimensionality was violated.

Table 4 reports the results for the dichotomous data fit with the GGUM. Results indicated that the fit indices examined here showed much higher power for detecting misfit in the GGUM than with the 2PL model. As with the results reported in Table 3, the Type I error rates were typically low for all of the indices examined here, except for Yen's Q_1 which had uniformly high Type I error rates under these conditions. However, the utility of these fit indices for detecting misfit when the GGUM was incorrectly fit to the data varied. Again, although Yen's Q_1 generally had high power for detecting this form of misfit, it also had high Type I error rates. Also consistent with the results presented in Table 3, the unadjusted chi-squares had lower power for detecting misfit except when the sample size was large and/or the data were generated from a multidimensional model. In contrast, the sample-size-adjusted chi-square values performed better than the unadjusted chi-squares. The SRMSR also performed well and tended to have the highest power for detecting misfit. The Type I error rates for the SRMSR were .00 and power was generally above .90 under most of the conditions simulated here. To illustrate these findings further, the distributions of the SRMSR and the adjusted chi-squares for doubles and triples are presented for representative conditions in Figures A1 and A2 in the online supplemental material. In contrast to the results of conditions in which the 2PL model was fit to the data, both the adjusted chi-squares for doubles and triples and the SRMSR had slightly lower power when the sample size was 250 and there were only 10 items. Nevertheless, these fit indices still provided moderate power and performed better than the other indices examined here.

In contrast to the results presented in Table 3, the results presented in Table 4 indicate that the PPP performed well when the underlying data were generated from the GGUM. Except when the sample size was small and there were few items, the Type I error rates were near .05 and the power was above .90. As such, although the PPP had difficulty identifying misfit of the 2PL model, this index appears useful for identifying misfit of the GGUM.

Tables 5 and 6 provide results for the polytomous data fit with the GRM and GGUM, respectively. The pattern of results largely replicated the findings in the dichotomous conditions. As shown in Table 5, when the GRM was fit to the data, Yen's Q_1 resulted in high Type I error rates and low power. In addition, the chi-squares for item doubles and triples as well as the SRMSR provided low Type I error rates and high power for detecting misfit when data were generated from the GGUM and fit with the GRM. In other words, these indices were useful for differentiating between dominance and ideal point models. However, consistent with the dichotomous conditions, both the adjusted chi-square values and the SRMSR had lower power for detecting misfit due to multidimensionality. When data were generated from a multidimensional GRM and fit with a unidimensional GRM, the chi-squares and SRMSR rarely identified misfit. Again, the PPP had low power for detecting misfit when the GRM was incorrectly fit to the ideal point data.

Table 6 presents the results for conditions in which the data were generated from a polytomous IRT model and fit with a polytomous GGUM. Here, both the SRMSR and the adjusted chi-squares for item singles, doubles, and triples performed relatively well. The Type I error rates under all conditions were .00. In addition, both the SRMSR and the adjusted chi-squares had high power for differentiating dominance and ideal point models. Although there was some variability in the unadjusted chi-square values, the power of the adjusted chi-squares was uniformly high under most conditions. The one exception was when the sample size ($N = 250$) and number of items in the measure ($i = 10$) were small, which resulted in slightly lower power for detecting misfit. This likely reflects the fact that IRT estimation requires larger sample sizes for estimation and suggests that results will become less accurate in samples as small as 250. Despite this limitation, it appears that the SRMSR and the chi-squares for item doubles and triples were some of the most useful indices for detecting misfit with polytomous data based on the conditions examined here.

Finally, Tables 7 and 8 show the results for the AIC and BIC with both dichotomous and polytomous data, respectively. Again, these results are presented separately because these fit indices are interpreted differently than the indices described above. Specifically, these are comparative fit indices rather than absolute fit indices and are compared across models to determine which model provides the best fit. As such, the results presented in Tables 7 and 8 indicate the percentages of the simulated data sets in which the fitted model provided the lowest values of AIC and BIC.

The results for these indices were consistent with the results for the adjusted chi-squares for doubles and triples and the SRMSR. When the data were generated from a dominance model (i.e., 2PL) and fit with the GGUM, the AIC and BIC identified misfit in 100% of the simulated samples. The opposite was also the case. When the data were generated from the GGUM and fit with a dominance model, misfit was identified in 100% of the cases. In other words, these indices were able to accurately identify the best fitting model when comparing dominance and ideal point models. Again, these indices only provide relative indicators of fit and, therefore, were not able to identify misfit in some cases. For example, when the data were generated from a 3PL model and fit with both the 2PL and GGUM models, the AIC and BIC suggested that the 2PL model fit the data better. This suggests a limitation of the AIC and BIC for detecting IRT model fit. These indices are not able to identify the *correct* model and can only accurately determine which model fits best. As such, these indices will be most useful when used to complement the adjusted chi-squares for doubles and triples and the SRMSR. In other words, the adjusted chi-squares and SRMSR can determine whether the

Table 5. Power and Type I Error Rates of the Model Fit Indices When Fit With the GRM in the Polytomous Conditions.

			Fitted Model (GRM)								
I	*N*	Data Gen. Model	Singles	Doubles	Triples	Singles*	Doubles*	Triples*	Yen's Q1	PPP	SRMSR
10	250	GRM	.00	.00	.00	.20	.00	.00	.41	.04	.20
		GGUM	.05	.03	.00	.46	1.00	1.00	.89	.20	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.18	.03	.00
	500	GRM	.00	.00	.00	.00	.00	.00	.64	.05	.00
		GGUM	.09	.18	.12	.27	1.00	1.00	.94	.24	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.34	.06	.00
	1,000	GRM	.00	.00	.00	.00	.00	.00	.86	.05	.00
		GGUM	.19	1.00	1.00	.33	1.00	1.00	.98	.54	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.66	.06	.00
	2,000	GRM	.00	.00	.00	.00	.00	.00	.95	.05	.00
		GGUM	.23	1.00	1.00	.27	1.00	1.00	1.00	.60	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.92	.05	.00
20	250	GRM	.00	.00	.00	.02	.00	.00	.27	.03	.00
		GGUM	.10	.38	.37	.20	1.00	1.00	.76	.28	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.16	.06	.00
	500	GRM	.00	.00	.00	.00	.00	.00	.33	.06	.00
		GGUM	.11	.55	.90	.29	1.00	1.00	.89	.24	.98
		MGRM	.00	.00	.00	.00	.00	.00	.21	.05	.00
	1,000	GRM	.00	.00	.00	.00	.00	.00	.49	.08	.00
		GGUM	.27	1.00	1.00	.37	1.00	1.00	.90	.63	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.29	.09	.00
	2,000	GRM	.00	.00	.00	.00	.00	.00	.73	.10	.00
		GGUM	.40	1.00	1.00	.45	1.00	1.00	.89	.78	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.45	.08	.00
40	250	GRM	.00	.00	.00	.00	.00	.00	.24	.02	.00
		GGUM	.22	1.00	1.00	.89	1.00	1.00	.69	.07	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.19	.06	.00
	500	GRM	.00	.00	.00	.01	.00	.00	.26	.05	.00
		GGUM	.24	1.00	1.00	1.00	1.00	1.00	.79	.34	.97
		MGRM	.00	.00	.00	.00	.00	.00	.20	.07	.00
	1,000	GRM	.00	.00	.00	.00	.00	.00	.29	.04	.00
		GGUM	.88	1.00	1.00	1.00	1.00	1.00	.91	.58	.99
		MGRM	.00	.00	.00	.00	.00	.00	.22	.06	.00
	2,000	GRM	.00	.00	.00	.00	.00	.00	.43	.04	.00
		GGUM	.97	1.00	1.00	1.00	1.00	1.00	.99	.80	1.00
		MGRM	.00	.00	.00	.00	.00	.00	.26	.04	.00

Note: *I* = number of items; *N* = sample size; Data Gen. Models = data generation models. Shaded cells indicate correctly specified models. Singles, doubles, and triples are Drasgow et al.'s (1995) chi-square model-data fit statistics.
*Indicates the sample-size-adjusted model fit index (Chernyshenko, Stark, Drasgow, & Roberts, 2007).

model fits well in an absolute sense and the AIC and BIC can be used to provide additional information about the best fitting model when multiple models are estimated.

Discussion

The present study examined the performance of a number of IRT model fit indices for detecting misfit in organizational research. Although past research has generally found positive results for the

Table 6. Power and Type I Error Rates of the Model Fit Indices When Fit With the GGUM in the Polytomous Conditions.

		Data Gen.	Fitted Model (GGUM)								
I	*N*	Model	Singles	Doubles	Triples	Singles*	Doubles*	Triples*	Yen's Q1	PPP	SRMSR
10	250	GRM	.59	.34	.07	.86	.79	.73	.97	.97	.96
		GGUM	.00	.00	.00	.00	.00	.00	.34	.03	.01
		MGRM	.30	.27	.14	.47	.40	.38	.79	.47	.35
	500	GRM	.63	.44	.28	.85	.75	.68	.97	.99	.85
		GGUM	.00	.00	.00	.00	.00	.00	.56	.04	.00
		MGRM	.31	.25	.20	.45	.36	.34	.75	.53	.29
	1,000	GRM	.63	.48	.33	.77	.67	.57	.97	.99	.80
		GGUM	.00	.00	.00	.00	.00	.00	.81	.03	.00
		MGRM	.46	.27	.23	.87	.63	.49	.77	.79	.24
	2,000	GRM	.80	.69	.45	.87	.80	.57	.92	1.00	.76
		GGUM	.00	.00	.00	.00	.00	.00	.89	.06	.00
		MGRM	.89	.60	.29	.95	.85	.52	.77	.95	.23
20	250	GRM	.60	.37	.15	.90	.83	.80	.95	.91	.79
		GGUM	.00	.00	.00	.00	.00	.00	.23	.04	.00
		MGRM	.74	.68	.66	.96	.88	.88	.66	.80	.71
	500	GRM	.81	.59	.34	.99	.98	.88	1.00	.98	.74
		GGUM	.00	.00	.00	.00	.00	.00	.26	.06	.00
		MGRM	.85	.86	.83	1.00	1.00	.96	1.00	.94	.83
	1,000	GRM	.89	.78	.51	1.00	.99	.90	1.00	1.00	.85
		GGUM	.00	.00	.00	.00	.00	.00	.37	.05	.00
		MGRM	1.00	.95	.92	1.00	1.00	1.00	1.00	.98	.99
	2,000	GRM	1.00	.95	.54	1.00	1.00	.94	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.42	.07	.00
		MGRM	1.00	1.00	.97	1.00	1.00	1.00	1.00	1.00	1.00
40	250	GRM	.58	.48	.37	.90	.80	.81	.87	.84	.56
		GGUM	.00	.00	.00	.00	.00	.00	.20	.03	.00
		MGRM	.51	.42	.42	1.00	.89	.88	.68	.61	.42
	500	GRM	.88	.60	.41	1.00	.96	.90	1.00	1.00	.88
		GGUM	.00	.00	.00	.00	.00	.00	.23	.05	.00
		MGRM	.96	.94	.90	1.00	1.00	.99	1.00	1.00	.94
	1,000	GRM	.91	.84	.59	1.00	1.00	.95	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.27	.04	.00
		MGRM	1.00	.98	.98	1.00	1.00	1.00	1.00	1.00	1.00
	2,000	GRM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		GGUM	.00	.00	.00	.00	.00	.00	.31	.04	.00
		MGRM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note: *I* = number of items; *N* = sample size; Data Gen. Models = data generation models. Shaded cells indicate correctly specified models. Singles, doubles, and triples are Drasgow et al.'s (1995) chi-square model-data fit statistics.
*Indicates the sample-size-adjusted model fit index (Chernyshenko, Stark, Drasgow, & Roberts, 2007).

indices examined here, these indices have never been compared under these conditions. In addition, the performance of many of these indices has not been examined under conditions that test their utility for differentiating between dominance and ideal point models. The present study helps to address this issue and demonstrates that the performance of these indices for detecting the types of misfit examined here can vary widely across conditions.

One contribution of this study was to examine misfit due to applying dominance models to ideal point data and vice versa. Results indicated that the indices examined here were generally good

Table 7. Power and Type I Error Rates of the AIC and BIC in the Dichotomous Conditions.

			Fitted Model			
			2PL		GGUM	
Item	N	Data Gen. Model	AIC	BIC	AIC	BIC
10	250	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	500	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	1,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	2,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
20	250	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	500	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	1,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	2,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
40	250	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	500	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	1,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00
	2,000	2PL	1.00	1.00	.00	.00
		3PL	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		M2PL	1.00	1.00	.00	.00

Table 8. Power and Type I Error Rates of the AIC and BIC in the Polytomous Conditions.

			Fitted Model			
			GRM		GGUM	
Item	N	Data Gen. Model	AIC	BIC	AIC	BIC
10	250	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	500	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	1,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	2,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
20	250	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	500	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	1,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	2,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
40	250	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	500	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	1,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00
	2,000	GRM	1.00	1.00	.00	.00
		GGUM	.00	.00	1.00	1.00
		MGRM	1.00	1.00	.00	.00

indicators of this type of major misfit. When data were generated from an ideal point model and fit with a dominance model, the power to detect misfit for many of the indices examined here was high. Similarly, when an ideal point model was fit to data generated from a dominance model, power was also high for detecting misfit. These findings are important because dominance and ideal point models represent substantial differences in the item response processes. As noted above, dominance models are the most widely used class of models in the organizational literature. However, recent research has suggested that ideal point models may be useful for *any* construct measures that require individuals to think about and report on how they typically think, feel, or behave (Drasgow et al., 2010). In addition, a growing body of research suggests that ideal point models may be useful for a number of constructs examined in the organizational literature such as employee personality (Carter et al., 2014; Stark et al., 2006), vocational interests (Tay et al., 2009), person-organization fit (Chernyshenko et al., 2009), performance ratings (Borman et al., 2001), and job attitudes (Carter & Dalal, 2010). Given that all of these constructs, and many more that might fit the ideal point response process, have traditionally been examined using dominance models, it appears that the type of misfit that can be adequately detected by several of the fit indices examined here is common in the organizational literature. In addition, this type of major misfit is likely to become more frequent as ideal point models are examined with additional constructs in the future.

One area where this type of misfit may be prevalent is in the person-environment (P-E) fit literature (Chernyshenko et al., 2009). Conceptually, the measurement of P-E fit should follow an ideal point response process. Individuals are only likely to indicate that they fit with a job if the job matches their level of the latent trait on some characteristic of interest. Although some initial work on this topic has been done (Chernyshenko et al., 2009), it is possible that the response process could vary for different types (e.g., person-organization, person-job, person-group fit) or sources (e.g., values, personality, interests) of fit and more research is needed to understand the implications of dominance and ideal point models for P-E fit research. This is just one example but the results of the present study can help to inform future research applying IRT to organizational constructs and comparing both dominance and ideal point models.

Across all conditions comparing dominance and ideal point models, the SRMSR and the adjusted chi-squares for item doubles and triples were the most accurate indicators of misfit. These indices generally had low Type I error rates and high power under many of the conditions examined here. Based on these results, we recommend estimating and interpreting both of these indices to evaluate the fit of IRT models. As done in the present study, the chi-squares for doubles and triples can be averaged across items and the means of these indices can be divided by their degrees of freedom to provide a scale-level index comparable to the SRMSR, which is already calculated at the scale level. Once this is done, chi-square/df ratios greater than 3 (Chernyshenko et al., 2001) and values of SRMSR greater than .05 (Maydeu-Olivares & Joe, 2014) are indicative of misfit. To help facilitate the use of these indices, R code for calculating the SRMSR and an Excel file for calculating the adjusted chi-squares for item doubles and triples are available at the following website: https://psychology.psy.msu.edu/pers_nye/IRT_syntax/.

The results of the present study also indicated that the AIC and BIC performed well under these conditions. However, it is important to remember that the AIC and BIC are comparative fit indices that only provide useful information when comparing the fit of multiple IRT models. In contrast to the chi-square fit statistics and the SRMSR, these indices do not provide information about absolute fit. Therefore, these indices will be most useful when comparing several plausible models.

Given the characteristics of the AIC, BIC, SRMSR, and the adjusted chi-squares, these indices should be interpreted together to provide the most comprehensive evaluation of IRT model fit. Although the AIC and BIC are relative fit indices, the SRMSR and adjusted chi-squares are absolute fit indices that can identify misfit regardless of which other models are tested. Nevertheless, the AIC and BIC identified the correct model when comparing ideal point and dominance models 100% of the time. Therefore, these indices can provide useful information under conditions in which other indices may be slightly less accurate. To combine these indices and effectively examine IRT model fit, multiple models will need to be tested and the AIC and BIC can be examined to identify the best fitting model out of those that are estimated. Once the best fitting model is identified, the SRMSR and the adjusted chi-squares for doubles and triples can be examined to determine absolute fit. If these indices suggest that the model does not fit the data well, the researcher will need to reevaluate the set of IRT models that are being tested to identify one that is more appropriate for the measure that is being used. This approach of using multiple fit indices to determine model fit is consistent with recommended practices in the SEM literature (Hu & Bentler, 1999; Kline, 2005). Past research has demonstrated that SEM fit indices may be sensitive to different types of model misfit and, therefore, reporting multiple indices provides a more comprehensive estimate of fit (Hu & Bentler, 1998). The results of the present study suggest that this is also the case in the IRT literature and that combining the fit indices examined here can help researchers to identify an appropriate model for the data and use IRT more effectively to advance organizational research.

A Note of Caution About the Use of IRT Fit Indices

Despite the positive results found when comparing dominance and ideal point models, the findings of the present study also suggest caution when comparing different dominance models and examining model fit. This is because the results presented here showed that many IRT fit indices are insensitive to differences between dominance models. For both dichotomous and polytomous data, power was near .00 under most conditions when either the 2PL model or the GRM, respectively, were fit to data generated from a different dominance model. Despite the positive results for the chi-square fit indices and the SRMSR when comparing dominance and ideal point models, even these indices were problematic when the incorrect dominance model was fit to the data. In other words, these indices were insensitive to situations in which too few parameters were estimated. This is important because it indicates that even the best indices examined here may be difficult to use to detect minor differences in model specification. Therefore, researchers should be cautious when interpreting model fit under these conditions and more research is needed to identify model-data fit indices that can detect misfit when an inappropriate dominance model is fit to the data.

The low power to differentiate between dominance models also applied to violations of the assumption of unidimensionality. When either the 2PL model or the GRM were fit to data generated from a multidimensional model, virtually all of the indices examined here had low power for detecting this form of misfit. This result is particularly surprising for the chi-squares for item doubles and triples. Past research has differentiated between first- (i.e., single items) and second-order (i.e., item doubles) fit statistics (Suárez-Falcón & Glas, 2003). Conceptually, first-order statistics should be most appropriate for identifying differences in predicted responses while second-order statistics have been designed to detect violations of unidimensionality. Research has tended to support this distinction (Glas, 1988; Yen, 1984). Therefore, it is surprising that the chi-squares for item doubles and triples exhibited such low power for detecting this form of misspecification. More research is needed to explore this issue further. Although this finding contradicts previous research on IRT model fit, it is not a serious impediment to the use of IRT. Instead, these results suggest that the dimensionality of the measure used for the IRT analyses needs to be addressed prior to IRT estimation. As noted above, this can be done using standard CFA software and techniques to verify that the measure is either unidimensional or multidimensional. After this has been determined, an appropriate IRT model can be selected given the dimensionality of the measure.

Overall, despite the mixed results, the findings of the present study provide an important contribution to the understanding of fit in IRT research. Again, although examining the fit of SEM is common practice, many articles that use IRT do not report fit indices (Foster et al., 2017). This is problematic because the advantages of IRT (described above and summarized in Table 1) can only be realized when the model actually fits the data. Consequently, it should be standard practice to report fit indices in the IRT literature as well. This is particularly salient when comparing dominance to ideal point models because past research has indicated that applying the incorrect model to the data can have substantial effects on the conclusions that are drawn from these analyses (Carter et al., 2014; Chernyshenko et al., 2001; Chernyshenko et al., 2007; Roberts et al., 1999). In addition, although the results of the present study contradict previous research suggesting that fit indices can detect multidimensionality (Glas, 1988; Suárez-Falcón & Glas, 2003; Yen, 1984), this assumption can be tested in other ways (e.g., by explicitly testing the dimensionality of the measure using CFA).

Similar alternatives are not available for differentiating dominance and ideal point models, making the assessment of fit that much more important for comparing these two classes of models. Therefore, despite the insensitivity of the adjusted chi-squares and SRMSR to minor differences in dominance models, these indices will be useful for differentiating dominance and ideal point models in future organizational research.

Conclusion

Although the results presented here indicate that IRT model-data fit indices are useful for detecting misfit under many conditions, they also illustrate the limitations of these indices. The general conclusion seems to be that these indices are useful for detecting major forms of misfit but are less successful at detecting minor differences between the assumptions of the model and the data. However, this type of major misfit is likely to be common in the organizational literature and to increase in importance as research continues to identify constructs that are represented best by ideal point models. Nevertheless, future research should examine the issue of fit further to understand the implications of these results. For example, although past research has shown that fitting dominance models to ideal point data can influence IRT results (Chernyshenko et al., 2007; Roberts et al., 1999), fitting a 2PL model to 3PL (i.e., two dominance models) data may not have a substantial influence on the estimation process. More research is needed to understand these effects and to identify IRT fit indices that are sensitive to this type of misfit.

Authors' Note

An earlier version of this article was presented at the annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL, April 2017.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Bo Zhang https://orcid.org/0000-0002-6730-7336

Supplemental material

Supplemental material for this article is available online.

Notes

1. Although both are based on Thurstone's work, a distinction can be made between ideal point models (Thurstone, 1928) and pairwise preference models (Thurstone, 1927). Specifically, Thurstonian ideal point models provide a way of scaling items and were first proposed in Thurstone's 1928 article. In contrast, Thurstone proposed pairwise preference models in a separate article published in 1927 as a way to model responses to forced-choice items (i.e., items where the respondent must choose from two or more statements). Here, pairwise preference models can be used with either ideal point or dominance response processes (Brown & Maydeu-Olivares, 2011; Maydeu-Olivares & Brown, 2010). We do not address issues related to forced-choice measures in the present study but ideal point models are examined.
2. In this equation, $P_{i1}(\theta) = 1 - \frac{1}{1+e^{-a_i(\theta-b_{i1})}}$ and $P_{iC}(\theta) = P_{iC-1}(\theta)$.

3. Traditional CFA analyses assume a dominance response process and, therefore, are difficult to use on scales developed from an ideal point perspective (Chernyshenko et al., 2007). Alternative approaches have been suggested (Davison, 1977; Roberts, Donoghue, & Laughlin, 2000), but more research is needed to examine ways of conducting factor analyses on ideal point data. A full discussion of this issue is beyond the scope of the present study, but interested readers are referred to Chernyshenko et al. (2007) and Habing, Finch, and Roberts (2005) for more details on these issues.
4. Maydeu-Olivares and Joe (2014) also developed a root mean square error of approximation (RMSEA). However, they noted that "it is best to use the SRMSR as a goodness-of-fit index. As a goodness-of-fit index, the SRMSR can be easily computed for models of any size. In particular, it can be computed for models with so many response patterns that the RMSEA$_2$ cannot be computed" (p. 319). In addition, Maydeu-Olivares and Joe also noted that the two indices were strongly related ($R^2 \geq .95$). Therefore, we focus on the SRMSR here.
5. The Ox code for all simulations is available in the online supplemental material.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akadémiai Kiadó.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171-187.

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

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*, 113-141.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965-973.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460-502.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI Work Satisfaction scale. *Personality and Individual Differences*, *49*, 743-748.

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. *Journal of Applied Psychology*, *99*, 564-586.

Carter, N. T., Dalal, D. K., Lake, C. J., Lin, B. C., & Zickar, M. J. (2011). Using mixed-model item response theory to analyze organizational survey responses: An illustration using the job descriptive index. *Organizational Research Methods*, *14*, 116-146.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology*, *84*, 610-619.

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

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*, 88-106.

Chernyshenko, O. S., Stark, S., & Williams, A. (2009). Latent trait theory approach to measuring person-organization fit: Conceptual rationale and empirical evaluation. *International Journal of Testing*, *9*, 358-380.

Dalal, D. K., & Carter, N. T. (2015). Consequences of ignoring ideal point items for applied decisions and criterion-related validity estimates. *Journal of Business and Psychology*, *30*, 483-498.

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

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education*, *5*, 17-34.

de La Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30*, 216-232.

DeMars, C. E. (2004). Type I error rates for generalized graded unfolding model fit indices. *Applied Psychological Measurement*, *28*, 48-71.

DeMars, C. E. (2005, August). *Scoring subscales using multidimensional item response theory models*. Poster presented at the annual meeting of the American Psychology Association, Washington, DC.

Doornik, J. A. (2009). An object-oriented matrix programming language Ox 6. Computer software.

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

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, *11*, 59-79.

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

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363-373.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to Support Selection and Classification Decisions* (Tech. Rep. 1311). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of item response theory practices in organizational research: Lessons learned and paths forward. *Organizational Research Methods*, *20*, 465-486.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*, 525-546.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology*, *48*, 82-98.

Habing, B., Finch, H., & Roberts, J. (2005). A Q-sub-3 statistic for unfolding item response theory models: Assessment of unidimensionality with two factors and simple structure. *Applied Psychological Measurement*, *29*, 457-471.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *3*, 424-453.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99-114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*, 828-845.

Joo, S. H., Lee, P., & Stark, S. (2017). Evaluating anchor-item designs for concurrent calibration with the GGUM. *Applied Psychological Measurement*, *41*, 83-96.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*, 499-518.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, *25*, 39-52.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*, 399-419.

Kim, J. S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*, 38-51.

Kline, R. B. (2005). *Principles and practices of structural equation modeling*. New York, NY: Guilford.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3-21.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, *45*, 935-974.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305-328.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*, 728-743.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437-455.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107-135.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. 6th ed. Los Angeles, CA: Muthén & Muthén.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*, 966-980.

Nye, C. D., Newman, D. A., & Joseph, D. L. (2010). Never say "always"?: Extreme item wording effects on scalar invariance and item response curves. *Organizational Research Methods*, *13*, 806-830.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517-529.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*, 207-230.

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

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

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

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.

Roznowski, M. (1989). Examination of the measurement properties of the Job Descriptive Index with experimental items. *Journal of Applied Psychology*, *74*, 805-814.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 18). Iowa City, IA: Psychometric Society.

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

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429-449.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583-639.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, *86*, 943-953.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, *89*, 497-508.

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*, 463-487.

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.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*, 138-152.

Stark, S., Chernyshenko, O. S., Nye, C. D., Drasgow, F., & White, L. A. (2017). *Moderators of the Tailored Adaptive Personality Assessment System Validity* (Tech. Rep. 1357). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Suárez-Falcón, J. C., & Glas, C. A. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143.

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.

Tay, L., & Drasgow, F. (2012). Adjusting the adjusted $\chi 2/\mathrm{df}$ ratio statistic for dichotomous item response theory analyses: Does the model fit?. *Educational and Psychological Measurement*, *72*, 510-528.

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.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*, 3-46.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*, 273-286.

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Wang, W. (2014). *A Bayesian Markov chain Monte Carlo approach to the generalized graded unfolding model estimation: The future of non-cognitive measurement* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125-145.

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychological Measurement*, *72*, 774-799.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*, 168-190.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84*, 551-563.

Author Biographies

Christopher D. Nye received his PhD from the University of Illinois in Urbana-Champaign and is currently an assistant professor of organizational psychology at Michigan State University. His research primarily involves personnel selection and assessment, individual differences, and organizational research methods.

Seang-Hwane Joo received his PhD from the University of South Florida and is a postdoctoral research fellow at KU Leuven, Belgium. His research focuses on item response theory, psychometrics, and multilevel modeling.

Bo Zhang is a doctoral candidate in the Industrial-Organizational Psychology program at the University of Illinois at Urbana-Champaign. His research focuses on psychometrics, organizational research methods, and applying them to address issues in personnel selection and the assessment of individual differences.

Stephen Stark received his PhD from the University of Illinois in Urbana-Champaign and is currently a professor and director of the Industrial-Organizational Psychology program at University of South Florida. His research focuses on item response theory methods and applications, noncognitive testing, selection, and organizational research methods.