Original Research Article

How Accurate and Consistent Are Score-Based Assessment Decisions? A Procedure Using the Linear Factor Model

Assessment
1–11

DOI: 10.1177/10731911221113568
journals.sagepub.com/home/asm
SAGE

Oscar Gonzalez[1], A. R. Georgeson[2], and William E. Pelham III[3]

Abstract

When scales or tests are used to make decisions about individuals (e.g., to identify which adults should be assessed for psychiatric disorders), it is crucial that these decisions be accurate and consistent. However, it is not obvious how to assess accuracy and consistency when the scale was administered only once to a given sample and the true condition based on the latent variable is unknown. This article describes a method based on the linear factor model for evaluating the accuracy and consistency of scale-based decisions using data from a single administration of the scale. We illustrate the procedure and provide R code that investigators can use to apply the method in their own data. Finally, in a simulation study, we evaluate how the method performs when applied to discrete (vs. continuous) items, a practice that is common in published literature. The results suggest that the method is generally robust when applied to discrete items.

Keywords

linear factor model, classification consistency, classification accuracy

In the social sciences, education, and medicine, tests and scales are often used for selection, such as to find the best candidates for a job, identify students with a minimum level of achievement, or determine which respondents need further psychological assessment. This article describes methods that apply to any selection scenario, but we frame our discussion using the example of screening measures: short assessments used to identify respondents who may have psychiatric disorders. Typically, scale item responses are summed (X^*; e.g., Achenbach & Rescorla, 2000; Allison et al., 2012), and then, a decision is made by comparing a respondent's summed score X^* to a cutpoint, X_c^* (Pepe, 2003). When scores are used to make decisions about individuals, it is critical to ensure that the selection process is accurate and consistent (American Educational Research Association [AERA] et al., 2014).

Classification accuracy (CA) refers to the probability of correctly assigning a respondent to the correct group. CA is measured as the agreement between the decision based on the summed scores and a reference class, representing the true condition of the respondent (e.g., a decision determined by a gold standard or true scores; Lee, 2010). Accurate classification supports the valid use of tests for decision-making (Lathrop, 2015). For example, suppose that a clinician administers the AQ10 (Autism Spectrum Quotient; Allison et al., 2012), and the respondent is referred for further diagnostic assessment for autism spectrum disorders if their score is above a cutpoint. The AQ10 would have high CA if it correctly distinguishes individuals who are likely to be diagnosed with autism spectrum disorder from those who are not.

Beyond accuracy, classification consistency (CC) refers to the probability that a respondent would receive the same classification across repeated administrations of the measure (Gonzalez et al., 2021; Lee, 2010; Livingston & Lewis, 1995). The concept of CC is similar to the test–retest reliability of a classification (Haertel, 2006; Lathrop, 2015). For the example above, the AQ10 would yield inconsistent classifications if a respondent is above the cutpoint today, but they would have been below the cutpoint if they would have received the assessment tomorrow. Note that CC assumes that there is no change in the respondents' level of

[1]The University of North Carolina at Chapel Hill, USA
[2]Arizona State University, Tempe, USA
[3]University of California, San Diego, USA

Corresponding Author:
Oscar Gonzalez, The University of North Carolina at Chapel Hill, Davie Hall, Campus Box 3270, Chapel Hill, NC 27559, USA.
Email: ogonza13@unc.edu

the construct (e.g., via maturation, practice effects, carry-over effects, or treatment effects) between the administrations of the measure—the only factor that affects the change in classification is measurement error (Gonzalez et al., 2021). There are many ways to estimate CA and CC from measures (Deng, 2011), but in this article, we focus on estimates based on latent variable models.

Recently, model-based estimates using item response theory (IRT) have been used to provide estimates of CA and CC in educational and psychiatric settings (e.g., Gonzalez et al., 2021; Gonzalez & Pelham, 2021; Lathrop & Cheng, 2013; Lee, 2010). These IRT-based methods fall short for three reasons. First, many researchers (e.g., clinical psychologists) are more familiar with the linear factor model than IRT models. Reise and Waller (2009) note that IRT models are the exception instead of the norm to analyze clinical assessments, although this is changing. Second, although most scales in the social sciences have discrete responses, some items may have continuous response scales, which are incompatible with IRT models. Examples of scales with continuous response scales include items whose response format is a continuous line segment, a visual analog scale, or time to complete task (Mellenbergh, 2017). Also, items that have many response categories (Thissen et al., 1983), such as those found in the European Social Survey (Davidov et al., 2008), may be treated as continuous. Third, for several reasons, investigators may prefer to analyze items with discrete response scales as though they are continuous, using the linear factor model (Beauducel & Herzberg, 2006; Jorgensen & Johnson, 2022; Li, 2016; Muthén & Kaplan, 1985; Rhemtulla et al., 2012). For instance, they may be unfamiliar with IRT, have limited sample sizes for accurate parameter estimation, or prefer a simpler model (e.g., the linear factor model has three parameters per item, and the number of parameters per item for IRT models depends on the number of response categories; Thissen, 2017). In the area of screening, examples of widely-used scales that have discrete item responses but are often treated as continuous include the Center of Epidemiologic Studies–Depression (CES-D) scale (e.g., Carleton et al., 2013) and the K6 scales (e.g., Bessaha, 2017). Thus, to fully realize the benefit of IRT-based advances for estimating CA and CC, these methods must be extended to the linear factor analysis framework.

There has been limited work on the estimation of CA in situations in which item responses are treated as continuous. Examples include Millsap and Kwok (2004), who showed how one can translate the parameters of a linear factor model into estimates of sensitivity and specificity to describe how well a measure classifies respondents, and Lai et al. (2017), who facilitated the implementation of these procedures with R code. In these two cases, however, the estimation of CC or the estimation of both CA and CC at specific levels of the latent construct were not discussed. Peng and Subkoviak (1980) also developed an approach to estimate CC that makes similar assumptions to our procedure, but they do not estimate CA or CC at specific levels of the latent construct. As such, there are two main contributions of this article. First, we extend the IRT-based procedure to estimate CA and CC (Gonzalez & Pelham, 2021; Lee, 2010) to handle responses that are treated as continuous. Second, we investigate how treating discrete items as continuous (i.e., analyzing discrete items with the linear factor model) affects the estimation of CA and CC. These contributions are important because (1) they would facilitate and promote the estimation of CA and CC for users of the linear factor model and (2) they would help determine whether researchers who routinely fit linear factor models to discrete data can still obtain a rough approximation of CA and CC.

Present Study

The purpose of this study is to provide applied researchers with tools, based on the linear factor model, to estimate CA and CC when they use scales for decision-making. First, we define four indices of model-based CA and CC and explain how these indices are estimated. Then, we illustrate our method by applying it to published data on the K6 screener for psychological distress (Kessler et al., 2003). Finally, we conduct a Monte Carlo simulation study to examine how CA and CC are affected when discrete items are analyzed as continuous. In the Supplemental Materials, we present R functions to conduct the proposed procedure with the linear factor model and a brief tutorial on how to use them.

Indices of CA and CC

Consider a hypothetical group of individuals who have identical latent variable scores η. At the item level, random measurement error would result in different observed item responses for these individuals, which in turn results in a distribution of possible sum scores X^* for each level of η. We refer to this conditional summed score distribution as $P(X^*|\eta)$ and some examples are depicted in the right panel of Figure 1. To estimate CA and CC, one needs $P(X^*|\eta)$, which can be determined from an IRT model or a linear factor model. In the appendix, we explain how one can use estimates of the factor loadings, intercepts, and residual variances (e.g., item parameters) to determine $P(X^*|\eta)$. The main takeaway of the appendix is that the linear factor model depends on the assumption that $P(X^*|\eta)$ is normally distributed, while that distribution can be non-normal for IRT models. Therefore, the performance of our procedure depends on meeting this assumption.

Using $P(X^*|\eta)$, one can define four indices (Gonzalez et al., 2021; Lee, 2010):

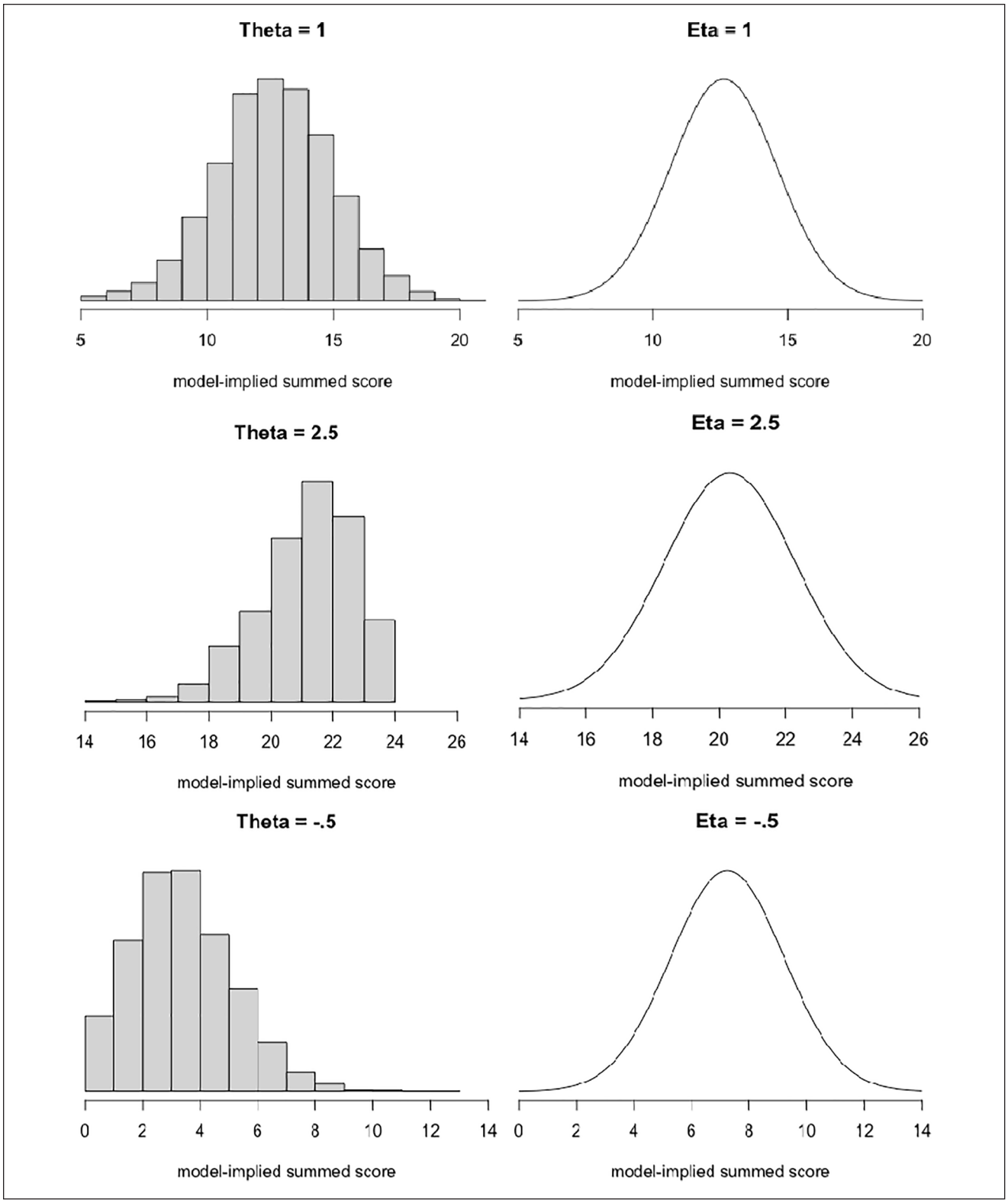

Figure 1. Left Panels: Summed Score Distribution Conditional on the Latent Variable (e.g., $P(X^* \mid \theta)$) From an Item Response Theory Model. Right Panels: Summed Score Distribution Conditional on the Latent Variable From a Linear Factor Model (e.g., $P(X^*|\eta)$). Note That the Range of the Scores is From 0 to 24.

1. *Conditional CA* (*CCA*). The probability of making a correct decision based on X^* at a specific value of η.
2. *Conditional CC (CCC)*. The probability making the same decision based on X^* across two parallel administrations of the measure at a specific value of η.
3. *Marginal CA (MCA)*. Weighted average of CCA estimates across the range of η.
4. *Marginal CC (MCC)*. Weighted average of CCA estimates across the range of η.

These four estimates range from 0 to 1, and higher values are better. Note that the four indices are specific to cutpoint c on X^*, so that, different X_c^* will have different CA and CC values.

Conceptual Sketch of the Method and Relative Advantages

In this section, we provide a brief explanation of the procedure to estimate CA and CC, which is most effective if the user is looking at the top right panel of Figure 1 while reading.

1. Estimate $P(X^*|\eta)$, as in the top right panel of Figure 1, at one value of η.
2. For a cutpoint[1] X_c^*, estimate the proportion of $P(X^*|\eta)$ at or above X_c^*, p_1, and below X_c^*, p_2.
3. Estimate CCA by checking if the X^* is above or below X_c^*. If $X^* \geq X_c^*$, CCA is p_1, else CCA is p_2. To estimate CCC, add $p_1^2 + p_2^2$.
4. Repeat Steps 1 to 3 for many more η values. Typically, η is normally distributed, so that, one can use equally spaced values between −2 and 2. These values are known as quadrature points.
5. Estimate the MCA and MCC by taking a weighted average of the CCA and CCC from the step above. The weights come from the quadrature points in Step 4 based on the height of the normal distribution (i.e., Gaussian quadrature), which are used to approximate the integral over η.

As mentioned above, the respondent's η is considered fixed, so that, $P(X^*|\eta)$ quantifies the uncertainty of a respondent's X^* at each level of η due to measurement error. In situations in which η is not expected to change, $P(X^*|\eta)$ provides a range of X^* that we are likely to observe across repeated administrations. If model assumptions are met, this property facilitates the estimation of CC because a single administration of the measure provides hypothetical information on test–retest performance, saving resources, and reducing participant burden (Gonzalez et al., 2021; Lee, 2010).

Illustration: Application of the Method to Published Example

K6 Scale

The K6 is a screener used to study psychological distress in the population (Kessler et al., 2003). Psychological distress is defined by Drapeau et al. (2010) as a combination of depression and anxiety symptoms which indicate emotional ill-being. The K6 is a screener that assesses how frequently an individual experienced six symptoms in the past 30 days: sadness, nervousness, restlessness, hopelessness, worthlessness, and the feeling that everything was an effort. The response scale has five categories (0—none of the time, 4—all the time), and an observed score X^* is estimated by summing all the items. Observed scores $X^* \geq 13$ have been found to identify participants with moderate psychological distress (Kessler et al., 2003).

For this illustration, we borrow the item parameters from the reference group reported in the study by Sunderland and colleagues (2012, Table 3) on the K6 scales. Data for the Sunderland et al. (2012) study were from the Australian National Survey of Mental Health and Well-being, and the item parameters come from a sample ($N = 2{,}761$) in which respondents were between the ages of 16 and 34, and 49.4% were women. We simulated 10,000 responses using the K6 item parameters from the graded response model (GRM; Samejima, 1969), and a normally distributed latent variable score with a mean and variance of 1, $N(1,1)$.[2] It is important to note that the item thresholds were asymmetric (e.g., had a positive skew). Then, the data were analyzed with the linear factor model, the parameters were saved, and we estimated CA and CC using quadrature points between −2 and 2 in steps of .05 using the R functions presented in the Supplemental Materials. The purpose of the illustration is to demonstrate the estimation of CC and CA under the linear factor model approach and compare it to CA and CC reference values from an IRT model, which are the data-generating parameters (further described below). Previous findings suggest that when discrete item responses are analyzed with the linear factor model, the factor loadings are attenuated (Beauducel & Herzberg, 2006; Jorgensen & Johnson, 2022; Li, 2016; Rhemtulla et al., 2012). As such, we expect that those CA and CC estimates would be smaller than the estimates from the IRT model because the relation between each item and the latent variable is underestimated.

Item Response Estimates of CA and CC as Reference Values. For our analyses, we use the CA and CC estimates from the IRT model as reference values because the procedure based on the IRT model accounts for the discrete nature of the items. We know that the item response variables are in truth discrete because they were designed that way by the scale constructors. Thus, by definition, a model

parameterized to reflect the discrete response options is more accurate ("closer to the truth") than a simpler model enforcing the assumptions of the linear factor model across item responses, which are not often tenable (e.g., model residuals are normally distributed; Wirth & Edwards, 2007). For this reason, both the illustration and simulation study consider the values from the IRT model as the reference case and evaluate to what extent an approach based on a simpler model (i.e., a linear factor model) produces similar results. As mentioned above, the CA and CC estimation require the item parameters, the cutpoint X_c^*, and the quadrature points. Sample size only plays a role in the precision of the item parameters, which is impacted by sampling error. Therefore, we obtained the reference values from the IRT model by treating the K6 item parameters as population values, and the X_c^* and quadrature points as fixed, and we estimated CA and CC using the cacIRT R-package (Lathrop, 2015). Moreover, we simulated many responses (e.g., N=10,000), which are analyzed to obtain factor model estimates from the linear factor model with small sampling variability (i.e., small standard errors).

Results. The top panel of Figure 2 shows the relation between the latent variable score and the model-implied summed score under the IRT model and the linear factor model. Although the linear factor model provides a rough approximation to the relations found by the data-generating (item response) model, there areas in which the summed score implied by the linear factor model is higher or lower than the summed score implied b the item response model. In the simulated dataset, a K6 cut score of $X_c^* \geq 13$ selects roughly 19% of the respondents. The θ value (i.e., the latent variable in the IRT model, analogous to η) that yields a model-implied $X_c^* = 13$ is roughly $\theta_c = .97$, while the corresponding η value is $\eta_c = 1.06$. Recall that the K6 parameters from the IRT model are treated as population values, and the highest standard error for an estimated parameter from the linear factor model was .014. The middle panel of Figure 2 shows the curves of the CCA estimates, and the MCA from the IRT model was .929 and for the linear factor model was .930. The bottom panel of Figure 2 shows the curves of the CCC estimates, and the MCC estimate from the IRT model was .902 and for the linear factor model was .900. Furthermore, we also examined the CA and CC estimates for $X_c^* = 20$, which roughly capture 3% of the individuals. The MCA for the IRT model was .980, while for the factor model was .992. On the other hand, the MCC for the IRT model was .973, while for the factor model was .988. It is likely that the MCA and MCC estimates were similar across approaches for $X_c^* = 13$ because of the conditional summed score distributions in the regions around X_c^*—the expected mean relation is on the top of Figure 2, the Var $(X^*|\eta)$ has a constant value of 1.97, and the average Var $(X^*|\theta)$ for $\theta > 0$ is 2.18. For $X_c^* = 20$, the estimates might be similar because most of the respondents will be ruled out with $X_c^* = 20$ because it is close to maximum score on the K6 of 24, which in turn yields high CA and CC regardless of the shape of the conditional summed score distribution. In this example, the CCA and CCC estimates differed but the linear factor model provided a close approximation to the MCA and MCC estimates from the data-generating model. As such, researchers who fit a linear factor model to K6 data (e.g., Bessaha, 2017) can obtain an approximate estimate of CA and CC even when items are analyzed as continuous.

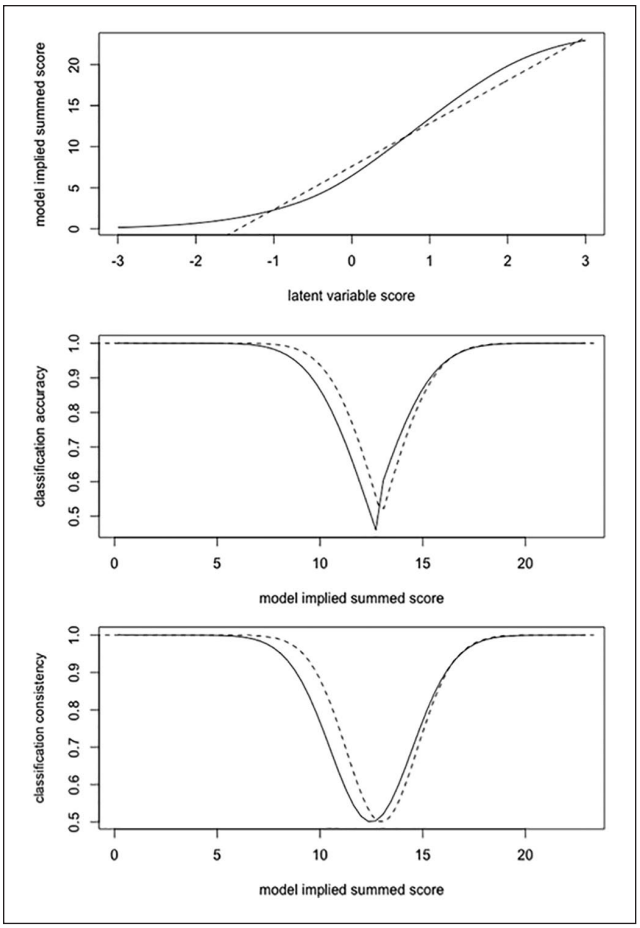

Figure 2. Top: Test Characteristic Curves for the Item Response Theory Model and the Linear Factor Model. Middle: Conditional Classification Accuracy Curves at Cutpoint 13 for the Item Response Theory Model and the Linear Factor Model. Bottom: Conditional Classification Consistency Curves at Cutpoint 13 for the Item Response Theory Model and the Linear Factor Model. For all Plots, Solid Lines Are for the Item Response Theory Model and Dashed Lines Are for the Linear Factor Model.

Simulation Study

The goal of the simulation is to determine whether the MCA and MCC estimates from data-generating models

with discrete items (Lee, 2010), MCA_D and MCC_D, are approximated by the MCA and MCC estimates from the linear factor model, MCA_C and MCC_C, at the population level. Recall that sampling variability does not affect our procedure per se (e.g., taking item parameters, determining conditional summed score distribution, imposing the cutpoint, and integrating results across η). Sampling error plays a role on the estimation of IRT model parameters or factor model parameters, which takes place prior to conducting the procedure. Implicitly, users of the procedure assume that there is precise estimation of item/factor model parameters.

To mimic common screening applications, discrete item responses to a unidimensional measure were generated, and IRT-based estimates are treated as the *reference values* as discussed above. Consistent with previous studies that factor loadings are underestimated when item response are discrete with a few categories (Beauducel & Herzberg, 2006; DiStefano, 2002; Jorgensen & Johnson, 2022; Li, 2016; Muthén & Kaplan, 1985; Rhemtulla et al., 2012; Thissen, 2017), we expect that as the number of items and number of response categories increase, the MCA_C and MCC_C estimates will be more similar to MCA_D and MCC_D.

Data Generation

Data were simulated using an ordered-categorical unidimensional factor model, similar to the simulations by Gonzalez and Pelham (2021). There are deterministic relations between the parameters from a categorical factor model and from the GRM (Wirth & Edwards, 2007), but we chose to simulate data from a categorical factor model because we believe that researchers would be more familiar with this metric. The factors varied were number of items (from 5 to 15), number of response categories (4, 5, 6, 7) per item (the lowest response category value was zero), and the distribution of the thresholds (symmetric or asymmetric). In total, there were 88 conditions, with $N = 20{,}000$ simulated cases generated per condition to mitigate the effect of sampling variability. The latent variable score was drawn from a standard normal distribution. The unique scores on each item were multivariate-normally distributed with means of zero, variances of one minus the communality of the item, and uncorrelated with each other and with the latent variable. The standardized item factor loadings per condition were equally spaced between 0.30 and 0.90, and the item thresholds were either symmetrically spaced from a standard normal distribution (i.e., evenly divided, with limits of −2.5 and 2.5) or asymmetrically spaced (i.e., the peak of the distribution fell to the left of the mean; moderately asymmetric condition), using the values provided by Rhemtulla et al (2012, see supplemental materials). After item responses were generated, a summed score for each respondent was computed, and the observed cutpoint X_c^* was half the maximum possible summed score in each condition (e.g., in a condition with seven items with five response categories, the maximum score is 28, so that, the cutscore was ≥ 14, selecting the top 50% of respondents), which corresponds roughly with the mean of the latent variable score. In the Supplemental Materials, we extend our simulation and present results for conditions that select the top 10% and top 25% of the respondents and also report MCC estimates using the Peng and Subkoviak (1980) procedure.

Data Analysis

The discrete item responses were analyzed with the linear factor model, parameters were saved, and the functions provided in the Supplemental Materials were used to estimate MCA_C and MCC_C. We used the cacIRT package with population item parameters to estimate the MCA_D and MCC_D. For both approaches, we used quadrature points between –2 and 2 in steps of .05 and normalized quadrature weights. The MCA_D and MCC_D were then compared to MCA_C and MCC_c using the tabled values and by estimating the root mean-squared difference (RMSD) and the relative mean difference (RMD), averaged across all conditions (see Tables S5–S10 in the Supplemental Materials for the raw difference and relative difference of estimates per condition). The RMSD was estimated by subtracting MCA_D and MCC_D from MCA_C and MCC_c, respectively, squaring the difference, averaging all values, and finally taking the square root. The RMSD value would indicate the mean absolute difference between estimates. The RMD was estimated by subtracting the MCA_D and MCC_D from MCA_C and MCC_c, respectively, then dividing by MCA_D and MCC_D. A positive RMD value would indicate that MCA_C and MCC_c overestimated MCA_D and MCC_D, and a negative RMD value would indicate that MCA_C and MCC_c underestimated MCA_D and MCC_D.

Results of Simulation

Tables 1 and 2 show the MCA_D, MCC_D, MCA_C, and MCC_c as a function of the number of items, number of response categories, and whether thresholds were symmetric or asymmetric. For the conditions with symmetric thresholds, the largest MCA_D and MCA_C difference was .008 (in the condition of six items with five response categories), the RMSD was .004, and the RMD was .004. Except for the conditions with four response categories, MCA_C slightly overestimated MCA_D, and as reflected by the positive RMSD and RMD. Furthermore, the largest MCC_D and MCC_C difference was .007 (in the condition of seven items with seven response categories), the RMSD was .004 and the RMD was .002, exhibiting similar patterns.

Table 1. Classification Accuracy Estimates for the Item Response Theory Model (in Regular Font) and the Linear Factor Model (in Bold Font) at a Cutpoint to Select the Top 50%.

		Symmetric thresholds				Asymmetric thresholds			
		Response categories				Response categories			
# Items	Approach	4	5	6	7	4	5	6	7
5	True discrete	0.807	0.810	0.816	0.816	0.899	0.889	0.883	0.873
	Continuous	**0.813**	**0.816**	**0.821**	**0.821**	**0.925**	**0.908**	**0.898**	**0.882**
6	True discrete	0.824	0.823	0.828	0.829	0.893	0.901	0.906	0.885
	Continuous	**0.823**	**0.831**	**0.831**	**0.833**	**0.911**	**0.916**	**0.921**	**0.893**
7	True discrete	0.834	0.834	0.838	0.840	0.893	0.910	0.922	0.894
	Continuous	**0.835**	**0.840**	**0.843**	**0.845**	**0.905**	**0.922**	**0.934**	**0.900**
8	True discrete	0.844	0.843	0.847	0.848	0.911	0.917	0.921	0.900
	Continuous	**0.844**	**0.847**	**0.850**	**0.853**	**0.925**	**0.928**	**0.932**	**0.905**
9	True discrete	0.849	0.851	0.854	0.856	0.925	0.922	0.932	0.906
	Continuous	**0.852**	**0.856**	**0.857**	**0.860**	**0.937**	**0.933**	**0.942**	**0.911**
10	True discrete	0.859	0.858	0.862	0.862	0.922	0.927	0.930	0.911
	Continuous	**0.858**	**0.862**	**0.865**	**0.866**	**0.933**	**0.937**	**0.939**	**0.915**
11	True discrete	0.863	0.864	0.867	0.868	0.921	0.931	0.930	0.916
	Continuous	**0.864**	**0.868**	**0.870**	**0.872**	**0.928**	**0.940**	**0.938**	**0.919**
12	True discrete	0.870	0.869	0.872	0.873	0.930	0.934	0.937	0.919
	Continuous	**0.868**	**0.873**	**0.874**	**0.877**	**0.939**	**0.941**	**0.945**	**0.923**
13	True discrete	0.874	0.874	0.877	0.878	0.938	0.937	0.937	0.923
	Continuous	**0.874**	**0.878**	**0.879**	**0.880**	**0.946**	**0.944**	**0.943**	**0.926**
14	True discrete	0.879	0.878	0.881	0.882	0.936	0.940	0.942	0.926
	Continuous	**0.878**	**0.880**	**0.884**	**0.885**	**0.943**	**0.947**	**0.948**	**0.929**
15	True discrete	0.881	0.881	0.885	0.885	0.942	0.942	0.947	0.928
	Continuous	**0.881**	**0.885**	**0.888**	**0.889**	**0.949**	**0.948**	**0.953**	**0.931**

For conditions with asymmetric thresholds, the largest MCA_D and MCA_C difference was .032 (in the condition of five items with four response categories), the RMSD was .010, and the RMD was .010. Across conditions, MCA_C slightly overestimated MCA_D. Furthermore, the largest MCC_D and MCC_C difference was .025 (in the condition of five items with four response categories), the RMSD was .013, and the RMD was .013, and patterns similar to MCA_C were observed. Across all conditions, there was not a clear pattern in the discrepancy across approaches, although this might be explained by the small differences between MCA_D and MCA_C, and between MCC_D and MCC_C across the board. Therefore, contrary to our hypotheses, ignoring the discrete nature of the items and fitting a linear factor model led to slightly larger, but similar MCA_C and MCC_C estimates compared to MCA_D and MCC_D, as reflected by the RMSD and RMD values.

In the Supplemental Materials, Tables S1 to S4 show the MCA_D, MCC_D, MCA_C, and MCC_c in conditions with X_c^* that select the top 25% and top 10% of respondents. Largely, we see that MCA_C and MCC_c overestimate MCA_D and MCC_D, although the largest deficit is less than 4%. Note that MCA_D, MCC_D, MCA_C, and MCC_c might be similar with these cutpoints because all the estimates are high regardless of the approach. In other words, most respondents are in the same class and are therefore more likely to be correctly and consistently classified. We expect to observe similar relations with very high or very low X_c^*, regardless of the shape of $P(X^*|\eta)$. Moreover, Figure S1 in the Supplemental Materials shows that MCC_c is higher than the CC estimates from the Peng and Subkoviak (1980) procedure. Across conditions, the correlation of the estimates across procedures ranged between $r = .95$ and .99.

Finally, we highlight two issues regarding cutpoints on X^*. Recall that CA and CC depend on X_c^*. For the simulation, the MCA_C and MCC_C estimates were higher in the conditions with asymmetric thresholds than the conditions with symmetric thresholds, but these values cannot be directly compared. The X_c^* was the same across both conditions, but the mass of the X^* distribution shifted to the left because the item thresholds were asymmetric. As such, X_c^* is located away from the mass of the distribution. Second, recall that the shapes of $P(X^*|\eta)$ for the linear factor model and the IRT model are slightly different (see Figure 1)—the conditional distribution for the linear factor model has a normal, smooth distribution that can take any value, while the conditional distribution for the IRT

Table 2. Classification Consistency Estimates for the Item Response Theory Model (in Regular Font) and the Linear Factor Model (in Bold Font) at a Cutpoint to Select the Top 50%.

		Symmetric thresholds				Asymmetric thresholds			
		Response categories				Response categories			
# Items	Approach	4	5	6	7	4	5	6	7
5	True discrete	0.744	0.746	0.754	0.750	0.860	0.845	0.836	0.822
	Continuous	**0.743**	**0.747**	**0.754**	**0.753**	**0.892**	**0.868**	**0.855**	**0.833**
6	True discrete	0.762	0.762	0.765	0.766	0.853	0.861	0.868	0.838
	Continuous	**0.756**	**0.766**	**0.766**	**0.770**	**0.873**	**0.881**	**0.887**	**0.848**
7	True discrete	0.780	0.775	0.777	0.779	0.850	0.873	0.890	0.850
	Continuous	**0.772**	**0.778**	**0.782**	**0.785**	**0.866**	**0.889**	**0.906**	**0.858**
8	True discrete	0.787	0.786	0.789	0.790	0.876	0.883	0.888	0.860
	Continuous	**0.783**	**0.788**	**0.791**	**0.795**	**0.893**	**0.898**	**0.903**	**0.866**
9	True discrete	0.796	0.796	0.799	0.800	0.895	0.890	0.903	0.869
	Continuous	**0.794**	**0.800**	**0.801**	**0.805**	**0.911**	**0.905**	**0.918**	**0.874**
10	True discrete	0.805	0.804	0.808	0.809	0.891	0.897	0.902	0.876
	Continuous	**0.803**	**0.808**	**0.812**	**0.814**	**0.904**	**0.910**	**0.913**	**0.880**
11	True discrete	0.814	0.812	0.816	0.816	0.889	0.902	0.901	0.882
	Continuous	**0.811**	**0.815**	**0.819**	**0.821**	**0.898**	**0.914**	**0.911**	**0.886**
12	True discrete	0.820	0.818	0.822	0.823	0.902	0.907	0.911	0.887
	Continuous	**0.816**	**0.823**	**0.824**	**0.829**	**0.913**	**0.917**	**0.922**	**0.891**
13	True discrete	0.826	0.824	0.829	0.829	0.913	0.911	0.911	0.892
	Continuous	**0.823**	**0.829**	**0.831**	**0.832**	**0.923**	**0.920**	**0.919**	**0.895**
14	True discrete	0.832	0.830	0.834	0.835	0.910	0.915	0.919	0.896
	Continuous	**0.829**	**0.832**	**0.837**	**0.838**	**0.919**	**0.924**	**0.926**	**0.900**
15	True discrete	0.837	0.835	0.838	0.839	0.919	0.918	0.925	0.900
	Continuous	**0.834**	**0.839**	**0.843**	**0.845**	**0.927**	**0.925**	**0.933**	**0.903**

model can be non-normal, is not smooth, and takes discrete values. Suppose that the cutpoint is at $X_c^* = 13$. Under the linear factor model, the next value higher than 13 with a given level of precision might be 13.0001 (or 13.00001, etc.), whereas for the IRT model, the next higher value is 14. The rounding when moving from a continuous model-implied X^* to a discrete model-implied X^* could introduce error that affects the precision of CA and CC, which in turn affects the comparisons. It is expected that the conditional distribution of X^* for the IRT model becomes smoother as the number of items and response categories increase.

Discussion

When tests and scales are used for decision-making, it is important to describe the decision process using estimates of CA and CC (AERA et al., 2014). This article introduced an analytical procedure to estimate CA and CC for the linear factor model. The proposed extension used the relations presented by Millsap and Kwok (2004) to develop a procedure to estimate CA and CC, similar to the analytical procedure by Lee (2010) and a simulation-based procedure similar to Gonzalez et al. (2021). Our proposed extension addresses a gap in the literature by enabling researchers who work with continuous item responses or who treat their item responses as continuous to estimate model-based CA and CC. In general, our proposed extension facilitates the estimation of CA and CC from a linear factor model, when before the estimation of CA and CC was only available for IRT models. Moving forward, researchers can use our procedure to report estimates of CA and CC for measures used for screening, selection, and decision-making. Also, we presented an illustration that researchers could replicate using the Supplemental Materials. The results from both the illustration and the simulation study suggest that when researchers treat discrete items as continuous, the CA and CC estimates from the linear factor model could slightly overestimate the CA and CC from the data-generating model where items are discrete. This difference would matter most in conditions with a few items (e.g., five or six items) and few response categories (e.g., four or five categories).

From our simulation results, we can provide general guidance for applied researchers using these methods. Whenever possible, we strongly recommend researchers to use latent variable models that match the response scale of the items. However, there might be instances in which researchers fit linear factor models to discrete items because that is the only latent variable model they know or because they do not have the sample size to estimate a model with more parameters. Simulation results suggest that model-based CA and CC from the linear factor model provide a reasonable approximation to the estimates from the data-generating model with discrete items if there are four response categories and threshold skewness is not too extreme. In other words, estimates of CA and CC from a linear factor model fit to discrete items are similar enough to provide a sense of the accuracy and consistency of the decision-making process.

The model-based estimates for CA and CC have several limitations. For example, our proposed procedure assumes that the measure is unidimensional and that the latent variable model fits the measure well. A future direction would be to extend this procedure to handle multiple factors and to study its performance in the presence of model misspecification or misfit (e.g., local dependence; Edwards et al., 2018). Furthermore, like other model-based estimates, the item parameters are treated as fixed for the estimation of CA and CC, but these item parameters are subject to sampling variability. When researchers use small sample sizes, the factor model parameters are not precise, and, our findings might not hold. A future direction would be to study how sampling variability and imprecise estimates of the factor model parameters or IRT model parameters impacts CA and CC. Finally, future research includes continuing to explore how CA and CC can quantify the impact of violations of measurement invariance on the selection process (Gonzalez et al., 2021; Gonzalez & Pelham, 2021; Lai et al., 2017; Millsap & Kwok, 2004). Hundreds of invariance studies have been published in psychology, but much of this work fails to clarify the extent to which the use of such scales is impacted by the items that exhibit bias (e.g., Nye et al., 2019). Furthermore, it is also unclear if the linear factor model can detect violations of invariance when fit to discrete items (e.g., Meade & Lautenschlager, 2004). As such, it would be important to investigate how the detection rate of noninvariance is reflected on the estimates of model-based CA and CC from linear models compared to IRT models. Overall, we encourage researchers to examine CA and CC in their measures using the methodology described in this article in tandem with other psychometric indices.

Appendix: Using the Linear Factor Model to Compute $P(X^*|\eta)$

Let X_{ij} be the observed score on item j of person i. When X_{ij} is continuous, the relationship between the observed score and the respondent's standing on the latent variable η_i could be described using the linear factor model:

$$X_{ij} = \tau_j + \lambda_j \eta_i + \Sigma_{ij}. \tag{1}$$

where τ is the item intercept, λ is the factor loading, and Σ_{ij} is the unique factor score for person i on item j. Recall that in many applied settings a summed score X^* is used for decision-making. We can use Equation 1 to derive two properties of X^* (Millsap & Kwok, 2004):

$$\begin{aligned} \mu_{X^*} &= \sum_j \tau_j + \left(\sum_j \lambda_j \right) \kappa, \ \mathrm{Var}(X^*) \\ &= \left(\sum_j \lambda_j \right)^2 \mathrm{Var}(\eta) + \sum_j \mathrm{Var}(\Sigma)_j . \end{aligned} \tag{2}$$

where μ_{x*} is the model-implied mean of X^*, Var (X*) is the model-implied variance of X^*, κ is the mean of η, and Var(η) is the variance of η.

Using the preceding developments, we can determine $P(X^*|\eta)$, the conditional summed score distribution. Two approaches have been previously studied to determine that distribution for IRT models: analytically using the approach by Lee (2010) or empirically using the approach by Gonzalez et al. (2021). Here, we focus on extending the analytical approach by Lee (2010) to the linear factor model, and we discuss the empirical approach by Gonzalez et al. (2021) in the Supplemental Materials—both approaches tend to produce similar results (Gonzalez et al., 2021). Millsap and Kwok (2004) indicated that $P(X^*|\eta)$ is normally distributed, so that, we can characterize $P(X^*|\eta)$ analytically using the conditional mean $E[X^*|\eta]$ and the conditional variance $\mathrm{Var}(X^*|\eta)$ of X^*. Given the relations in Equation 2, we can determine that

$$\begin{aligned} E\left[X^* | \eta \right] &= \sum_j \tau_j + \left(\sum_j \lambda_j \right) \eta, \ \mathrm{Var}(X^* \mid \eta) \\ &= \sum_j \mathrm{Var}(\Sigma_j). \end{aligned} \tag{3}$$

Note that Var($X^*|\eta$) does not vary as a function of η because Var(η) at a specific value of η is zero, which follows from the assumption of homogeneity of the residual variance. Note that $\sum_j \tau_j$ and $\sum_j \lambda_j$ are the intercept and slope of an unbounded line mapping η to X^*, and at each value of η, the X^* has a constant variance, $\sum_j \text{Var}(\Sigma)_j$. The η to X^* mapping is similar to the test characteristic curve (TCC) estimated for item responses models, with three exceptions: the TCC is bounded by the minimum and maximum values of X^*, can be S-shaped, and it does not have a constant variance at each level of the latent variable (Thissen, 2000). The right panel of Figure 1 shows examples of $P(X^*|\eta)$ from the linear factor model, which differs from the conditional summed score distributions obtained from an IRT model (Lee, 2010; left panel of Figure 1)—the former are continuous and normal at all levels of the latent variable while the latter are discrete and can take non-normal shapes. Thus, the performance of our procedure depends on how closely $P(X^*|\eta)$ adheres to a normal distribution at each level of η.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research was in part supported by NIH funding: DA053137 (A.R.G), AA030197 (W.E.P.), and DA055935 (W.E.P.).

ORCID iD

Oscar Gonzalez https://orcid.org/0000-0001-7122-8799

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The procedure defines the model-implied reference class for CA estimation as being at or above the η value that yields the model-implied X_c^*. For the example presented in the illustration section, if researchers use a K6 cutpoint of 13, then the model-implied reference class used by the procedure is defined by $\eta \geq 1.06$ because $\eta = 1.06$ yields a model-implied K6 cutpoint of 13 using Equation 3 in the appendix for the conditional mean. The X_c^* is typically determined empirically or by subject-matter experts. If the cutpoint were given in the η metric (e.g., cutpoint at $\eta = 1.5$), it could be transformed to the model-implied X^* using Equation 3.
2. We simulated θ to have a mean of 1 because some items had very extreme b-parameters (above 3). A consequence of this decision is that when we fit the factor model and constrain the θ distribution to have a mean of 0, the b-parameters are rescaled by subtracting 1 from their original values.

References

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms and profiles: An integrated system of multi-informant assessment*. University of Vermont Department of Psychiatry.

Allison, C., Auyeung, B., & Baron-Cohen, S. (2012). Toward brief "red flags" for autism screening: The short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*, 202–212.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Bessaha, M. L. (2017). Factor structure of the Kessler psychological distress scale (K6) among emerging adults. *Research on Social Work Practice*, *27*, 616–624.

Carleton, R. N., Thibodeau, M. A., Teale, M. J., Welch, P. G., Abrams, M. P., Robinson, T., & Asmundson, G. J. (2013). The Center for Epidemiologic Studies–Depression scale: A review with a theoretical and empirical examination of item content and factor structure. *PLOS ONE*, *8*(3), Article e58067.

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly*, *72*, 420–445.

Deng, N. (2011). *Evaluating IRT- and CTT- based method of estimating classification consistency and accuracy indices from single administrations* [Unpublished doctoral dissertation]. University of Massachusetts, Amherst.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*, 327–346.

Drapeau, A., Beaulieu-Prévost, D., Marchand, A., Boyer, R., Préville, M., & Kairouz, S. (2010). A life-course and time perspective on the construct validity of psychological distress in women and men. Measurement invariance of the K6 across gender. *BMC Medical Research Methodology*, *10*, 1–16.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*, 138–149.

Gonzalez, O., Georgeson, A. R., Pelham, W. E., III, & Fouladi, R. T. (2021). Estimating classification consistency of screening measures and quantifying the impact of measurement bias. *Psychological Assessment*, *37*, 596–609.

Gonzalez, O., & Pelham, W. E., III. (2021). When does differential item functioning matter for screening? A method for empirical evaluation. *Assessment*, *28*, 446–456.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Greenwood.

Jorgensen, T. D., & Johnson, A. R. (2022). How to derive expected values of structural equation model parameters when treating discrete data as continuous. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(4), 639–650.

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., Howes, M. J., Normand, S.-L. T., Manderscheid, R. W., Walters, E. E., & Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry*, *60*, 184–189.

Lai, M. H., Kwok, O. M., Yoon, M., & Hsiao, Y. Y. (2017). Understanding the impact of partial factorial invariance on selection accuracy: An R script. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 783–799.

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation, 20*, Article 18.

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, *37*, 226–241.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*, 1–17.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*, 936–949.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388.

Mellenbergh, J. G. (2017). Models for continuous responses. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 153–166). Chapman and Hall.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189.

Nye, C. D., Bradburn, J., Olenick, J., Bialko, C., & Drasgow, F. (2019). How big are my effects? Examining the magnitude of effect sizes in studies of measurement equivalence. *Organizational Research Methods*, *22*, 678–709.

Peng, C. Y. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement*, *17*, 359–368.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354–373.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 1–97.

Sunderland, M., Hobbs, M. J., Anderson, T. M., & Andrews, G. (2012). Psychological distress across the lifespan: Examining age-related item bias in the Kessler 6 Psychological Distress Scale. *International Psychogeriatrics*, *24*, 231–242.

Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159–183). Lawrence Erlbaum.

Thissen, D. (2017, January). Similar DIFs: Differential item functioning and factorial invariance for scales with seven ("plus or minus two") response alternatives [Paper presentation]. 81st International Meeting of the Psychometric Society, Asheville, NC, United States.

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, *7*, 211–226.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, 12, 58–79.