A Comparison of Limited-Information Test Statistics for a Response Style MIRT Model

Joshua Starr and Carl F. Falk

McGill University

Scott Monroe

University of Massachusetts, Amherst

David D. Vachon

McGill University

This is an Accepted Manuscript of an article published by Taylor & Francis in *Multivariate Behavioral Research* on October 26, 2020, available online: https://www.tandfonline.com/doi/full/10.1080/00273171.2020.1828024.

The authors would like to thank Li Cai for comments on an earlier draft of this manuscript; any errors in the manuscript are the responsibility of the authors. We acknowledge the support of the Natural Science and Engineering Research Council of Canada (NSERC), (funding reference number RGPIN-2018-05357 and DGECR-2018-00083). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence RGPIN-2018-05357]. Address all correspondence to: Carl F. Falk, Department of Psychology, McGill University, 2001 McGill College, 7th Floor, Montreal, QC H3A 1G1, Canada. Tel: 514.398.6133. Email: carl.falk@mcgill.ca

Abstract

An increased use of models for measuring response styles is apparent in recent years with the multidimensional nominal response model (MNRM) as one prominent example. Inclusion of latent constructs representing extreme (ERS) or midpoint response style (MRS) often improves model fit according to information criteria. However, a test of absolute model fit is often not reported even though it could comprise an important piece of validity evidence. Limited information test statistics are candidates for this task, including the *full* (M_2), *ordinal* (M_2^*), and *mixed* (C_2) statistics, which differ in whether additional collapsing of univariate or bivariate contingency tables is conducted. Such collapsing makes sense when item categories are ordinal, which may not hold under the MNRM. More generally, limited information test statistics have gone unevaluated under nominal data and non-ordinal latent trait models. We present a simulation study evaluating the performance of M_2, M_2^*, and C_2 with the MNRM. Manipulated conditions included sample size, presence and type of response style, and strength of item slopes on substantive and style dimensions. We found that M_2 sometimes had inflated Type I error rates, M_2^* always had little power, and C_2 lacked power under some conditions. M_2 and C_2 may provide complementary and valuable information regarding model fit.

Keywords: model fit, limited-information test statistics, Likert-type items, multidimensional item response theory, nominal response model, response styles

1. Introduction

One of the most popular ways to measure unobservable psychological constructs is to use a self-report questionnaire with Likert-type items. For example, participants may be asked to indicate levels of agreement to items on a 5-point scale (e.g. 0 = strongly disagree, 1 = disagree, 2 = neutral, 3 = agree, 4 = strongly agree). Examples of constructs measured in this way include social support (Maurer, Mitchell, & Barbeite, 2002), personality traits (Goldberg, 1992; Paulhus & Vazire, 2007), and self-esteem (Tafarodi & Swann, 2001). However, responses to such items are seldom reflective of only the constructs of interest. One possible source of extraneous influence is known as a *response style*. In brief, response styles represent a preference for certain response options, and this preference may be unrelated to the target construct. Common response styles include extreme responding (ERS), whereby individuals preferentially select the endpoints of the scale (e.g. 0 or 4), and midpoint responding (MRS), whereby the middle category is often selected (e.g. 2; Baumgartner & Steenkamp, 2001; Paulhus, 1991).

Although a number of sophisticated multidimensional item response theory (MIRT) models have recently emerged for response styles (e.g., Bockenholt, 2012, 2017; Jonas & Markon, 2018; Khorramdel & von Davier, 2014; Liu & Wang, 2019), important questions remain regarding model fit. In the present manuscript, we focus on the multidimensional nominal response model (MNRM) for response styles (Bolt & Newton, 2011; Falk & Cai, 2016; Falk & Ju, 2020). In our experience, use of the MNRM almost always results in better fit than standard models that omit style dimensions (e.g., generalized partial credit model; Muraki, 1992) according to information criteria such as AIC or BIC. However, statistical tests of overall model fit are rarely, if ever, reported for response style MIRT models. Traditional full-information test statistics such as Pearson's χ^2 and the likelihood ratio statistic G^2 exhibit poor performance with

IRT models in general, especially with long tests and/or polytomous items (i.e., more than 2 categories per item). These statistics perform poorly due in part to sparseness in multi-way contingency tables among the items. Limited information test statistics, such as M_2, M_2^*, and C_2 have been developed to address this problem (Cai & Hansen, 2013; Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2005, 2006;), and additional supplementary information such as RMSEA and TLI are computable based on these tests (e.g., Cai & Monroe, 2014; Maydeu-Olivares & Joe, 2014).[1] Such tests and accompanying information may comprise an important piece of validity evidence for latent variable models in general (Markus & Borsboom, 2013), and are potential candidates for reporting alongside response style MIRT models.

Unfortunately, there is a lack of recommendations regarding which limited information test statistic is most appropriate to use with response style MIRT models in general, and the MNRM in particular. To date, we are unaware of any previous research that has evaluated use of these test statistics based on either simulations or theoretical grounds with the response style MNRM, and even more generally with other response style models, the unconstrained nominal model or nominal data, or other non-ordinal latent trait models. Such comparisons are important because M_2 may not always be computationally efficient nor perform well with polytomous items, yet M_2^* and C_2 may further collapse categories within items in a way that does not make sense for certain MNRM variants where ordinality may not hold. To complicate matters, choice of a different test statistic may provide a different picture of overall model fit.

To illustrate this latter problem, consider results based on fitting several models with the MNRM to data ($N = 803$) from the Affective and Cognitive Measure of Empathy (ACME),

1. In addition to sparsity, all such statistics may require numerical integration, which is feasible for models with few latent traits, but still problematic for models with many latent traits.

developed by Vachon and Lynam (2016). The ACME consists of three subscales, Affective Resonance (AR), Affective Dissonance (AD), and Cognitive Empathy (CE), and each subscale is composed of twelve 5-point Likert-type items. Broadly conceived, AR is designed to measure empathic concern and general compassion, AD captures inappropriate affect – e.g., taking pleasure in others' pain – and CE is the ability to recognize and understand emotions in others. We followed a strategy that is not uncommon in test construction by examining each construct separately.[2] For brevity and as the pattern of results was similar, we only report results for the AR subscale, with results for AD and CE presented in the Supplementary Materials. All analyses were run in flexMIRT® (Cai, 2017), and details of the MNRM and limited information test statistics will be presented in a later section.

We established a baseline for model fit with an IRT model that includes only a dimension representing AR. We additionally considered the possibility that ERS (+ERS), MRS (+MRS), or both style constructs at once (+ERS+MRS) might be other plausible factors that could be added to improve fit. We fit all of these models with correlated factors and all items loading on substantive and style dimensions. As typically found by others (e.g., Falk & Cai, 2016; Falk & Ju, 2020), AIC showed improved model fit for both style factors (+ERS+MRS; Table 1). However, an inspection of M_2, M_2^*, and C_2, along with RMSEA and TLI, suggests that the test statistics disagree on the adequacy of model fit (Table 1). First, almost all models are rejected by the test statistics at the $\alpha = .05$ level. However, for the three +ERS models, M_2^* actually fails to reject the null. Second, M_2^* cannot be computed for models with both style constructs (+ERS+MRS) as there are negative degrees of freedom. Finally, RMSEA and TLI provide

2. We have retained all participants regardless of their standing on quality control checks. Our use of the subsequent models is mainly for illustrative purposes, and we do not make any firm substantive conclusions from the results of these analyses.

conflicting information across test statistics. M_2 does not yield TLI values indicating good fit ($>$ .90) until both ERS and MRS are included in the model. However, TLI based on C_2 looks better ($\geq$ .95) with only one style construct in the model, and TLI based on M_2^* looks good ($\geq$ .98) with just the baseline model. In contrast, RMSEA suggests a different pattern of results, with worse fit when computed from C_2 than from M_2 or M_2^*. RMSEA based on C_2 would suggest that there are more gains in fit if ERS is added than with MRS, yet the opposite sometimes holds true for RMSEA based on M_2 or M_2^*.

[Table 1 near here]

The takeaway here is that M_2, M_2^*, and C_2 offer different conclusions regarding model fit in the presence of style constructs. This problem is compounded by the fact that the researcher lacks crucial guidelines with which to make an informed choice to follow one test statistic over another, as there is currently a dearth of knowledge about how M_2, M_2^*, and C_2 behave with response style MIRT models, particularly when not all dimensions assume ordered categories. If such behavior were known, it may be easier to know whether each test statistic is sensitive to misspecification of style constructs, and what type of model modification may be most warranted, if any. Furthermore, the quality of TLI and RMSEA may be dependent on an initial evaluation of the test statistics themselves. Our goal for this paper is to therefore compare the relative performance of limited information test statistics with such models. In what follows, we more fully discuss the MNRM in Section 2. We explain limited-information fit in conceptual and technical detail in Section 3, along with a literature review of previous theoretical and simulation research. In Sections 4 and 5, we present the methods and results of a simulation study to evaluate test statistic behavior in the presence of response styles. To conclude, we will discuss the main findings with consideration to applied research.

2. Multidimensional Nominal Response Model (MNRM)

The MNRM (Thissen, Cai, & Bock, 2010; Thissen & Cai, 2016) is a divide-by-total model based in part on the unidimensional nominal response model by Bock (1972). Numerous authors have now used and introduced the model as being useful for measuring response styles (e.g., Bolt & Johnson, 2009; Bolt, Lu, & Kim, 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Falk & Ju, 2020; Johnson & Bolt, 2010; Ju & Falk, 2019; Kieruj & Moors, 2013). To formally introduce notation, suppose that an item is polytomous with K total categories indexed by $0,1,\ldots,K-1$. The traceline for the MNRM for category *k* of the item can be written as follows using scalar notation:

$$T(k|\boldsymbol{\theta}) = \frac{\exp(a_1 s_{k1}\theta_1 + a_2 s_{k2}\theta_2 + \cdots + a_D s_{kD}\theta_D + c_k)}{\sum_{m=0}^{K-1}\exp(a_1 s_{m1}\theta_1 + a_2 s_{m2}\theta_2 + \cdots + a_D s_{mD}\theta_D + c_m)} \quad (1)$$

where $\theta_1,\ldots,\theta_D$ represent *D* latent constructs of interests, $a_1,\ldots,a_D$ are slopes for each dimension, $s_{k1},\ldots,s_{kD}$ are scoring function values specific to category *k* and each dimension, and c_k is a category specific intercept. In some parameterizations of the model (Falk & Cai, 2016; Thissen et al., 2010; Thissen & Cai, 2016) and that implemented by flexMIRT® (Cai, 2017) and used in this manuscript, the intercepts are not directly estimated: $\boldsymbol{c} = \boldsymbol{T\gamma}$, where $\boldsymbol{c}$ is a *K*-length vector of intercepts, $\boldsymbol{T}$ is a $K \times (K-1)$ contrast matrix using a Fourier basis, and $\boldsymbol{\gamma}$ contains $K-1$ parameters. Note that with the exception of the latent traits, all of the aforementioned parameters may vary across items ($j = 1,\ldots,n$), whereas the values of the latent trait (if they were known) vary across study participants ($i = 1,\ldots,N$). That is, we have omitted such subscripts for notational simplicity. Maximum marginal likelihood (MML) estimation is often used for estimating item parameters (Bock & Lieberman, 1970; Bock & Aitkin, 1981), which requires distributional assumptions for the latent traits. Here we assume multivariate

normality, $\boldsymbol{\theta} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with a mean vector $\boldsymbol{\mu}$ (usually fixed to all zero for single group models), and covariance matrix, $\boldsymbol{\Sigma}$ (with the diagonal usually fixed to unity for identification).

We denote Y_{ij} a random variable for item *j* and participant *i*, and y_{ij} its observed response or realization. MIRT models estimated using MML in general can often be written in terms of the model-implied probability of a response pattern for all items, $\boldsymbol{y}_i = [y_{i1}\ y_{i2} \cdots y_{in}]$,

$$\pi_i = \pi(\boldsymbol{y}_i|\boldsymbol{\omega}) = \int f(\boldsymbol{y}_i|\boldsymbol{\omega}, \boldsymbol{\theta})\phi(\boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{\Sigma})d\boldsymbol{\theta} \quad (2)$$

where $\boldsymbol{\omega}$ is a vector of length v that includes all free model parameters (both item parameters and covariances among latent dimensions). $\phi(\cdot)$ is the density function for the latent traits (multivariate normal in our application), and $f(\boldsymbol{y}_i|\boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{j=1}^{n} P(Y_{ij} = y_{ij}|\boldsymbol{\theta}, \boldsymbol{\omega})$ is the conditional probability mass function for the response pattern, which represents a product of relevant tracelines since $P\left(Y_{ij} = y_{ij}\middle|\boldsymbol{\theta}, \boldsymbol{\omega}\right) = \prod_{k=0}^{K_j-1} T(k|\theta)^{1_k(y_{ij})}$ where $1_k(\cdot)$ is an indicator function equal to one when its input is equal to *k*, and zero otherwise. In other words, if the model is correct, π_i represents the probability of observing response pattern $\boldsymbol{y}_i$.

The version of the MNRM in Equation 1 is most similar to that presented by Falk and colleagues (Falk & Cai, 2016; Falk & Ju, 2020). Additional constraints are often imposed for both identification of the model, and for defining substantive and style constructs. As discussed by Falk and Ju (2020), the parameterization and identification constraints may vary depending on the software used to estimate the model. Here, we focus primarily on the parameterization of scoring function values and slopes as available in flexMIRT® (Cai, 2017) and *mirt* (Chalmers, 2012). In some recent applications of the MNRM (Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017), and the approach taken in this manuscript, we fix the scoring functions to prespecified values that make sense based on substantive theory. This strategy is most easily seen when examining scoring function values for a single latent construct, but across

all categories for a single item. For instance, if the categories for an item and latent construct have ordinal scoring functions for dimension *d* and a 5-category item, $[s_{0d}\ s_{1d}\ s_{2d}\ s_{3d}\ s_{4d}\ ] = [0\ \ 1\ \ 2\ \ 3\ \ 4]$, then the relationship between the construct and the item is equivalent to that modeled by the generalized partial credit model (GPCM; Muraki, 1992). Such scoring function values are typically congruent with the substantive construct of interest when categories are assumed ordinal; an increase (or decrease) in the latent trait tends to result in a choice of a higher (or lower) category. With ERS, the scoring function values may be fixed such that the endpoint categories indicate a response towards one end of the latent trait, and the middle categories indicate a response towards the other end of the latent trait: $[s_{0d}\ s_{1d}\ s_{2d}\ s_{3d}\ s_{4d}\ ] = [1\ \ 0\ \ 0\ \ 0\ \ 1]$. With MRS, a similar strategy is used, except the middle category has a "1" and the other categories have a zero scoring function value: $[s_{0d}\ s_{1d}\ s_{2d}\ s_{3d}\ s_{4d}\ ] = [0\ \ 0\ \ 1\ \ 0\ \ 0]$. It is also possible to estimate scoring function values (at least up until some identification constraints) as in the original nominal response model, so as to reveal how the categories are related to a latent construct. In such applications, for studying response styles it is often the case that a strong first dimension resembles the substantive trait, and a secondary dimension resembles ERS (Bolt & Johnson, 2010; Kieruj & Moors, 2013).

In the case of fixed scoring function values, if slopes are constrained equal across items (or fixed to 1 with the variance of the relevant latent construct estimated), then the model becomes analogous to a partial credit model (Masters, 1982). Some authors have used this strategy for measuring substantive and style constructs such as ERS and MRS (e.g., Bolt & Newton, 2011; Wetzel & Carstensen, 2017). Falk and Cai (2016) showed how it is possible to instead freely estimate slopes across items for both style and substantive factors, but retain fixed scoring function values. That is, all a_d that are of interest are freed parameters, yet all s_{kd} are

fixed to prespecified values depending on the dimension. It is even further possible in this latter setup to also estimate correlations among factors, even if all items load on all constructs, provided that scoring functions are not linearly dependent across latent dimensions. It is this latter approach that we used with the ACME data in the previous section to estimate models with a substantive trait only, and with combinations of ERS and MRS.

As such models are relatively new within the methodology literature, circulation and use has primarily focused there, with fewer applications outside of this literature (cf. Schneider, 2018; Stone, Schneider, Junghaenel, & Broderick, 2019). Potential applications have thus focused on adjustment of factor score estimates either on an observed score metric or latent metric (Bolt & Newton, 2011; Dowling, Bolt, Deng, & Li, 2016; Stone et al., 2019), the potential of style factors to distort across group item functioning (Bolt & Johnson, 2009), survey features that may elicit response styles (e.g., Kieruj & Moors, 2013), or applications to survey/study design (Adams, Bolt, Deng, Smith, & Baker, 2019). The potential consequences of fitting a misspecified MNRM in the context of response styles is thus difficult to ascertain due to a small research base, although previous simulation research has indicated that recovery of factor scores becomes slightly worse if relevant style factors are omitted (e.g., Falk & Cai, 2016), and differential item functioning may sometimes be obscured (e.g., Bolt & Johnson, 2009). Preliminary evidence thus far suggests that score estimates that are adjusted for style may have improved validity in some situations (e.g., Schneider, 2018).

Before moving on to limited information fit statistics in the context of the MNRM, we note that Tutz (2019) has argued that the partial credit version of the MNRM (implying also less restricted versions of the MNRM) do not represent *ordinal* latent trait models. This is most intuitively seen in how the scoring functions for constructs such as ERS and MRS are not in the

same order as the original categories (i.e., not monotonic), and do not match the order of the scoring function values for the main substantive dimension. This observation may become important when trying to form intuition about whether certain limited information fit statistics are more appropriate than others.

3. Limited Information Test Statistics

3.1 Overview of Limited-Information Testing

One approach to adjudicating overall model fit, adopted by full-information test statistics such as Pearson's X^2 and the likelihood ratio G^2 is to compare model-implied probabilities of for all possible response patterns to corresponding proportions observed in the actual data. For *n* polytomous items with K_j response categories for item *j*, let the model-implied multinomial response pattern probabilities under MML estimation be denoted $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \ldots, \hat{\pi}_C)'$, and the corresponding observed proportions be denoted $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_C)'$, where $C = \prod_{j=1}^{n} K_j$. Since these probabilities (or proportions) must sum to 1, there are $C - 1$ independent probabilities or proportions that could be used for model testing. Intuitively, a well-fitting model is one in which the model-implied probabilities for the response patterns closely match the observed proportions.

However, when there are many items, many categories, or both, accurately constructing a well-calibrated test statistic is challenging due to *sparseness* (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2006). Take for example any of the ACME subscales with 12 Likert-type items, each with five response options. In such a case, there are 5^{12} (or more than 244 million) cells in the full contingency table, which means nearly all of the cells (or observed proportions) remain empty assuming a realistic sample size N. In this situation, full-information test statistics typically have poorly-calibrated Type I error rates, and thus have limited utility for evaluating IRT models (Maydeu-Olivares & Joe, 2005). A related challenge is computational in nature:

computing all C of the probabilities for the full contingency table, as required for Pearson's X^2, can be impractical.

Limited-information statistics address these challenges by using lower-order probabilities and proportions (i.e., first-order, second-order, etc.), as lower-order marginal tables will necessarily be better-filled than the full multiway table. Thus, limited-information statistics are less impacted by sparseness. To date, the most popular limited-information statistics, such as M_2, use first- and second-order margins. An example of these sub-tables for two items from the ACME AR subscale (see Introduction Section) is presented in Table 2. As with full-information statistics, intuitively, a well-fitting model is one in which the model-implied probabilities closely match the observed proportions. Furthermore, in many practical settings, the number of first and second-order cells will be small compared to C. For example, for any of the ACME subscales, there are approximately 3,400 cells in the collection of first and second-order tables. As a result, limited-information statistics can be much less computationally demanding than full-information statistics.

[Table 2 near here]

The limited-information statistics M_2, M_2^*, and C_2 all use up to second-order marginal tables (hence, the subscript of 2), but differ in how the probabilities and proportions are compared. For M_2 (Maydeu-Olivares & Joe, 2005), a full set of mathematically independent first- and second-order model-implied probabilities is compared to the corresponding set of sample proportions. This set is obtained by excluding the probabilities with response categories of 0. For example, in Table 2, the shaded cells are not used to calculate M_2. Thus, M_2 relies on $q_1 = \sum_{j=1}^{n}(K_j - 1)$ mathematically independent first-order quantities and $q_2 = \sum_{j=1}^{n-1}\sum_{j\prime=j+1}^{n}\left(K_j - 1\right)\left(K_{j'} - 1\right)$ mathematically independent second-order quantities.

Though the lower-order tables used in M_2 will necessarily be better-filled than the full multiway table, the bivariate contingency tables may still exhibit sparseness with polytomous items. Cai and Hansen (2013) reasoned that when items are designed to measure the same primary construct, the responses are expected to have a positive relationship due to the influence of a common latent variable. So, it is expected that some observed response combinations will be common, while others may be rare. Returning to Table 2, the sample proportions are greater along the diagonal, and lower or even zero elsewhere, especially for response pairs that are inconsistent (e.g., category 4 on item 6, and category 1 on item 5).[3] Despite the sample size ($N = 803$), the observed frequency is less than five for nine of the 25 bivariate cells. This problem will be exacerbated when items have large loadings (or slopes) on the common latent trait, and/or items have many categories. In a simulation study, Cai and Hansen (2013) found that Type I error rates for M_2 can be too small when items are polytomous. Moreover, when the numbers of items and categories are large, M_2 can still be slow to compute.

A strategy to address this shortcoming is to collapse or summarize the lower-order tables used in M_2, thereby avoiding sparseness. For M_2^* (Cai & Hansen, 2013; see also Joe & Maydeu-Olivares, 2010), the first-order probabilities and proportions are summarized using item means as follows:

$$\hat{\mu}_j = \sum_{k=0}^{K_j-1} k\hat{\pi}_j^{(k)}; \;\; m_j = \sum_{k=0}^{K_j-1} kp_j^{(k)}, \tag{2}$$

where $\hat{\pi}_j^{(k)}$ and $p_j^{(k)}$ refer to the model-implied probability and observed proportion, respectively, for item j and category k (e.g., see Table 2). So, whereas M_2 relies on q_1

3. If items differ greatly in overall difficulty, these high frequencies may not cluster along the diagonal, but will still have a similar pattern where category pairs further toward at least one of the off-diagonals will have the lowest frequencies.

independent first-order probabilities, M_2^* relies on only $r_1 = n$ item means. For M_2^*, the second-order probabilities and proportions are similarly summarized, using the items scores to calculate item-pair moments as follows:

$$\hat{\mu}_{jj'} = \sum_{k=0}^{K_j-1} \sum_{k'=0}^{K_{j'}-1} kk' \hat{\pi}_{jj'}^{(kk')}; \; m_{jj'} = \sum_{k=0}^{K_j-1} \sum_{k'=0}^{K_{j'}-1} kk' p_{jj'}^{(kk')}, \tag{3}$$

where $\hat{\pi}_{jj'}^{(kk')}$ and $p_{jj'}^{(kk')}$ are the model-implied probability and observed proportion, respectively, for item j with category k, and item j' with category k' (e.g., see Table 2). Whereas M_2 relies on q_2 independent second-order probabilities, M_2^* relies on $r_2 = n(n-1)/2$ item-pair moments. As an example, Table 3 presents the item means and item-pair moment for the two example items in Table 2. Also, note that when the numbers of items and categories are large, M_2^* will be faster to compute than M_2.

[Table 3 near here]

Another statistic that avoids the potential sparseness in the bivariate contingency tables for polytomous items is C_2 (Cai & Monroe, 2014; Monroe & Cai, 2015). This statistic can be considered a mixed, or hybrid, version of M_2 and M_2^*. Like M_2^*, C_2 uses the r_2 item-pair moments to avoid sparseness. However, Cai and Monroe (2014) reasoned that sparseness rarely affects univariate tables, and therefore summarizing these tables using item means, as done for M_2^*, is likely unnecessary and could result in lower power. Instead, C_2 relies on the q_1 independent first-order probabilities, just like M_2. An advantage of C_2 over M_2^* is that for some combinations of numbers of items and categories, the latter statistic does not have positive degrees of freedom, and cannot be used for testing. For a summary of information regarding these statistics and information available to each, see Table 4.

[Table 4 near here]

3.2 Definitions of Limited-Information Test Statistics

Next, we present equations for the family of limited-information test statistics (Joe & Maydeu-Olivares, 2010), and present M_2, M_2^*, and C_2 as special cases of this family. Let $\mathbf{L}_\eta$ be an $s \times C$ reduction matrix that defines how the full contingency table is collapsed; that is, $\mathbf{L}_\eta$ defines the particular limited-information statistic. For examples of $\mathbf{L}_\eta$ leading to M_2 and M_2^*, see Cai and Hansen (pp. 253-257, 2013). Then, let $\hat{\boldsymbol{\eta}} = \mathbf{L}_\eta \hat{\boldsymbol{\pi}}$ be the s-length vector of summaries of the model-implied probabilities (e.g., the independent first- and second-order probabilities) and let $\boldsymbol{h} = \mathbf{L}_\eta \boldsymbol{p}$ be the corresponding sample statistics. The limited-information quadratic form test statistic is

$$Q_\eta = N(\boldsymbol{h} - \hat{\boldsymbol{\eta}})' \hat{\boldsymbol{\Upsilon}}_\eta (\boldsymbol{h} - \hat{\boldsymbol{\eta}}). \tag{4}$$

with weight matrix

$$\boldsymbol{\Upsilon}_\eta = \boldsymbol{\Xi}_\eta^{-1} - \boldsymbol{\Xi}_\eta^{-1} \boldsymbol{\Delta}_\eta \left(\boldsymbol{\Delta}_\eta' \boldsymbol{\Xi}_\eta^{-1} \boldsymbol{\Delta}_\eta \right)^{-1} \boldsymbol{\Delta}_\eta' \boldsymbol{\Xi}_\eta^{-1} \tag{5}$$

evaluated at the MML estimates $\hat{\boldsymbol{\omega}}$. In this last equation, $\boldsymbol{\Xi}_\eta$ denotes N times the asymptotic covariance matrix of $\boldsymbol{h}$, and $\boldsymbol{\Delta}_\eta$ denotes the matrix of derivatives of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\omega}$. If two conditions related to $\mathbf{L}_\eta$ are satisfied,[4] then the asymptotic null distribution of Q_η is χ^2 with $s - v$ degrees of freedom.

For M_2, let $\boldsymbol{\eta}_1$ be the vector of all linearly independent first- and second-order model-implied probabilities, and let $\boldsymbol{h}_1$ be the vector of corresponding sample statistics. In this case, $s = q_1 + q_2$, and $\mathbf{L}_{\eta_1}$ is defined such that $\boldsymbol{\eta}_1 = \mathbf{L}_{\eta_1} \boldsymbol{\pi}$ and $\boldsymbol{h}_1 = \mathbf{L}_{\eta_1} \boldsymbol{p}$. Then, M_2 is defined as

$$M_2 = N(\boldsymbol{h}_1 - \hat{\boldsymbol{\eta}}_1)' \hat{\boldsymbol{\Upsilon}}_{\eta_1} (\boldsymbol{h}_1 - \hat{\boldsymbol{\eta}}_1), \tag{6}$$

4. To establish the asymptotic distribution for the family, it is necessary that: 1) $\mathbf{L}_\eta$ has full row rank, s, and $\mathbf{1}_C'$ is not in its row span; and 2) $\boldsymbol{\Delta}_\eta$ has full column rank, v (p. 396, Joe & Maydeu-Olivares, 2010).

with $\widehat{\boldsymbol{\Upsilon}}_{\eta_1}$ calculated as in Equation 5, using $\mathbf{L}_{\eta_1}$. The degrees of freedom for M_2 is $q_1 + q_2 - v$.

For M_2^*, let $\boldsymbol{\eta}_2$ be the vector of all item-means and item-pair moments, and let $\boldsymbol{h}_2$ be the vector of corresponding sample statistics. In this case, $s = r_1 + r_2$, and $\mathbf{L}_{\eta_2}$ is defined such that $\boldsymbol{\eta}_2 = \mathbf{L}_{\eta_2}\boldsymbol{\pi}$ and $\boldsymbol{h}_2 = \mathbf{L}_{\eta_2}\boldsymbol{p}$. Then, M_2^* is defined as

$$M_2^* = N(\boldsymbol{h_2} - \widehat{\boldsymbol{\eta}}_{\boldsymbol{2}})'\widehat{\boldsymbol{\Upsilon}}_{\eta_2}(\boldsymbol{h_2} - \widehat{\boldsymbol{\eta}}_{\boldsymbol{2}}), \tag{7}$$

with $\widehat{\boldsymbol{\Upsilon}}_{\eta_2}$ calculated as in Equation 5, using $\mathbf{L}_{\eta_2}$. The degrees of freedom for M_2^* is $r_1 + r_2 - v$.

Finally, for C_2, let $\boldsymbol{\eta}_3$ be the vector of all linearly independent first-order probabilities and item-pair moments, and let $\boldsymbol{h}_3$ be the vector of corresponding sample statistics. In this case, $s = q_1 + r_2$, and $\mathbf{L}_{\eta_3}$ is defined such that $\boldsymbol{\eta}_3 = \mathbf{L}_{\eta_3}\boldsymbol{\pi}$ and $\boldsymbol{h}_3 = \mathbf{L}_{\eta_3}\boldsymbol{p}$. Then, C_2 is defined as

$$C_2 = N(\boldsymbol{h}_3 - \widehat{\boldsymbol{\eta}}_3)'\widehat{\boldsymbol{\Upsilon}}_{\eta_3}(\boldsymbol{h}_3 - \widehat{\boldsymbol{\eta}}_3), \tag{8}$$

with $\widehat{\boldsymbol{\Upsilon}}_{\eta_3}$ calculated as in Equation 5, using $\mathbf{L}_{\eta_3}$. The degrees of freedom for C_2 is $q_1 + r_2 - v$.

3.3 Previous Studies Comparing M_2, M_2^*, and C_2

There are very few previous simulation studies that have examined all three of the limited information test statistics simultaneously. In fact, we are aware of only a comparison of M_2 and M_2^* (Cai & Hansen, 2013), an evaluation of C_2 albeit using a different estimation approach (Monroe & Cai, 2015), and a small study comparing all three by Cai and Monroe (2014). In the latter study, the three test statistics were evaluated with a 4, 6, or 8 item test at a single sample size ($N = 500$). Unidimensional graded response models (Samejima, 1969) were fit to data from a unidimensional true model in order to study Type I error rates, and to data from the same model with an additional small substantive factor in order to study power. Across these previous studies, M_2 tended to have inaccurate Type I error rates (often lower than expected) – a trend that appears to exacerbate with more items. C_2 and M_2^* tended to maintain better Type I error rates, and C_2 has been shown to have higher power than both M_2 and M_2^*. In addition, sometimes M_2^*

has nonpositive degrees of freedom with questionnaires with few items, limiting its general utility and suggesting that perhaps the first- and second-order margins have been collapsed too far.

Note that the weights (indices k and k' in the summations in Equations 2 and 3) that reduce first- and second-order elements are ordinal and the same as the scoring function values for substantive dimensions under the MNRM as described in Section 2. Equivalently, these are also the scoring function values under the GPCM, which has been described as an ordinal latent trait model (Tutz, 2019). Indeed, in describing M_2^*, Maydeu-Olivares (2013) repeatedly refers to the test statistic as one for ordinal data: "… it only makes sense to use means and cross-products, and, therefore M_{ord}, when the item categories are ordered (hence its name)" (p. 81; where M_{ord} is notation for M_2^*). C_2 makes similar assumptions when reducing second-order information. We may then wonder about the appropriateness of C_2 and M_2^* for response style MIRT models that have been argued to not be ordinal, and in which the order of the categories (and scoring function) differs for constructs such as ERS and MRS. There is nothing wrong with the underlying statistical theory in reducing first- or second-order information and such test statistics would be expected to maintain nominal Type I error rates (Joe & Maydeu-Olivares, 2010), yet previous research with M_2^* and C_2 indicates that sometimes there is diminished power if too much information is lost due to collapsing (Cai & Monroe, 2014).

To our knowledge, relative test statistic performance remains unknown and untested on response style MIRT models. But this issue seems important given our review of the literature and the realization that some of the collapsing under C_2 and M_2^* may result in a loss of information that may adversely affect the performance of these statistics. In what follows, we conduct a Monte Carlo simulation study to evaluate the utility of limited-information test

statistics across varying conditions – sample size, strength of loadings, presence or absence of response styles – and offer researchers some clarity about which test statistic is most appropriate when modeling substantive constructs and response style traits simultaneously, in the context of Likert-scale measurement tools.

4. Method

Data were generated from one of three models: 1) unidimensional GPCM (a single substantive dimension; Model 1); 2) bidimensional model with substantive dimension (GPCM) and ERS (Model 2); and 3) bidimensional model with GPCM and MRS (Model 3). For each bidimensional model, we assumed uncorrelated substantive and response style factors, as previous investigations have often found these to be weakly related (e.g., Falk & Cai, 2016). The size of slope parameters was manipulated on a per-dimension basis such that all weak/strong and substantive/style combinations were satisfied. This resulted in two additional conditions for Model 1 (weak, strong) and four conditions for each of Models 2 and 3 (weak/weak, weak/strong, strong/weak, strong/strong). The idea behind varying slope strengths was twofold. First, we predicted that strong loadings on the substantive dimension may result in more sparseness in the bivariate contingency tables, as stronger loadings imply increased inter-item correlation and thus second-order marginal frequencies that cluster more closely around the bivariate diagonal. Second, stronger loadings on the response style factor may result in categories that are more severely out of order than would otherwise be true with weaker slopes. Each generating model produced data for $N = 500$ and $N = 1000$ subjects, and included twelve 5-category items across all conditions. In total then, there were then 20 unique data generation conditions, and we simulated 1000 replications per condition.

Item parameters for data generation were chosen in accordance with published item parameters in the MNRM context (Falk & Cai, 2016) and parameter estimates extracted from real Likert-scale data (Vachon & Lynam, 2016). Our objective here was to select a set of parameters that offers good coverage of the substantive latent trait and some slope and intercept variability across items. To avoid items so extreme that the marginal probability of any given response fell too low (defined as .05 or smaller) or had unrealistic looking response functions, we changed intercept parameters slightly when the substantive dimension had strong versus weak slopes. For the substantive dimension, half of the items had a slope of 1 and the remaining 1.2, whereas with weak slopes these values were .55 or .75.[5] For style factors (ERS or MRS), items were split into four groups with .45, .60, .55, or .65 for weak slope conditions and 1.2, 1.35, 1.3, or 1.4 for strong slope conditions. See Table 5 for an example of the strong/strong condition with Models 2 and 3, and Supplementary Materials for the remaining data-generating parameters.

[Table 5 near here]

All three types of data-generating models were crossed with the same three types of fitted models. When combined with the 20 data generating conditions, this yields 60 different unique cells under which to evaluate the performance of the test statistics, or 60,000 fitted models. When the true model is fit to the data, we evaluate Type I error rates. Technically, fitting Models 2 or 3 to data from Model 1 also evaluates Type I error rates as Model 1 is nested within these other two models. When the unidimensional GPCM (Model 1) is fit to either of the

5. Despite what appear to be slightly small standardized loadings for these items (e.g., Table 5; computed using flexMIRT®), the strong and weak conditions for the unidimensional model and a standard normal latent trait correspond to marginal reliabilities of .91 and .84, respectively. The information provided for style factors is expected to be less since the effective number of categories per item is fewer (e.g., see also Falk & Ju, 2020).

bidimensional models (Models 2 or 3), we can evaluate power to detect omitted style factors. Also to evaluate power with an incorrectly specified style factor, we fit Model 2 to data generated with Model 3, and vice versa. Scoring functions were fixed to prespecified values for all dimensions under true and estimated models and are the same as mentioned earlier in this manuscript: substantive factor [0 1 2 3 4], ERS [1 0 0 0 1], and MRS [0 0 1 0 0]. Correlations between constructs were fixed to zero.

All models were fit using flexMIRT® (Cai, 2017) with rectangular quadrature with 49 equally spaced points between -6 and 6 for each latent dimension. Models were estimated using maximum marginal likelihood with the Expectation-Maximization algorithm (EM-MML; Bock & Aitkin, 1981). To strike a balance between computational time and precision of convergence, we set the maximum number of E-step iterations to 2000 and the convergence criteria for the E-step and M-step to 1×10^{-4} and 1×10^{-9}, respectively.

Our theoretical expectations were as follows. With larger sample sizes, Type I error under the null distribution will approach nominal rates, and power under the alternative will improve. Under the null, a well calibrated test statistic should be approximately chi-square distributed with a mean equal to its degrees of freedom (*df*), and variance equal to twice the *df* (Agresti, 1996). This is likely not always the case for M_2 (Cai & Hansen, 2013; Cai & Monroe, 2014). We believe that M_2 may exhibit poor performance in terms of both Type I error and power when substantive loadings are strong, as this condition is likely to induce more sparseness in the second-order margins. In turn, M_2^* may perform well when substantive loadings are strong. However, M_2^* may have little power to detect a misspecified model as the ordinality assumption about the categories is not met with data from Models 2 and 3. Finally, C_2 may provide a good

compromise among the three limited-information fit measures as bivariate sparseness is reduced due to collapsing, yet some information from univariate margins is retained.

5. Results

5.1 Type I Error

All but 6 fitted models converged in under 2,000 iterations and results from models that did not converge were omitted from our results. In what follows, we present Type I error results involving only Models 1 and 2, as results with Model 3 – both as a data generating model and fitted model – were similar (see Supplementary Materials for full results). While we follow Cai and Monroe (2014) and report Type I error rates at three different alpha levels (α = .01, .05, .10), we only discuss accuracy at α = .05 because patterns are similar across all levels. When the true model was fit to the data and substantive slopes were *weak*, accurate rejection rates persisted across all three fit statistics and both samples sizes. For example, Type I error ranged from .036 to .06 under Model 1 (Table 6), and between .035 and .063 under Model 2 (Table 7). When substantive slopes were *strong*, M_2 exhibited inflated rejection rates. For example, at N = 500 rejection rates were as high as .10 or .11 under all strong substantive slope conditions. Rejection rates improved with a larger sample, but still ranged from .068 to .087. In contrast, M_2^* and C_2 maintained better Type I error rates (between .041 and .062) at both sample size conditions and when substantive slopes were strong. A comparison of the empirical means and variances of M_2, M_2^*, and C_2 reveals that the distribution of M_2 appears to have greater variance than the chi-square reference distribution, and this was especially true at N = 500 with strong substantive slopes, where it tended to be greater than 3 times the degrees of freedom. When fitting Models 2 or 3 to data from Model 1, a similar pattern of results was observed as already described, with

one exception. In particular, Type I error rates for M_2 tended to fall below nominal rates (.024 and .026) with weak substantive dimensions and at N = 1,000 (See Supplementary Materials).

[Table 6 near here]

[Table 7 near here]

5.2 Power

For models investigating power, all but 3 converged in under 2,000 iterations. We discuss two types of misspecification starting with an omitted style factor when Model 1 (a unidimensional GPCM) was fit to data from Model 2 (Table 8) or Model 3 (Table 9), which represent bidimensional models with ERS and MRS, respectively. Overall, M_2 had the highest power to detect model misspecification and this pattern held across *all* studied conditions. But, recall that since M_2 is less well-calibrated under the null, added care should be taken when evaluating M_2 for power. It stands out that M_2^* exhibited extremely low power across these conditions (ranging from .017 to .085), indicating an unsatisfactory capacity to flag misspecification. Although it might be expected that power to detect misspecification would increase with sample size and when style slopes are strong – indicating a greater misspecification – this pattern only held for M_2 and C_2. For example, considering only conditions where substantive slopes are strong, rejection rates of C_2 were .130 (N = 500) and .213 (N = 1000) when ERS slopes were weak, as compared to .986 (N = 500) and 1.000 (N = 1000) when ERS slopes were strong.

Given weak power by M_2^*, we focus more closely on the relative performance of C_2 and M_2. C_2 had better power than M_2^*, but not as good as M_2. For example, in the case of weak substantive and weak ERS slopes with N = 1000 observations, the rejection rate of M_2 (.830) at the α = .05 level was almost eight times greater than that of C_2 (.112). In the case of weak

substantive and weak MRS slopes with N = 1000 subjects, a similar relationship was observed. When response style slopes were strong, M_2 again exhibited higher power than C_2, but the gap between the test statistics was smaller. In only one such case (weak substantive slopes, strong MRS slopes, and N = 500 subjects) was the rejection rate of M_2 (1.000) more than two times greater than that of C_2 (.307). The power of C_2 improved quickly as sample size doubled. In fact, given larger sample sizes, weaker substantive slopes, and stronger ERS slopes, C_2 (.945) had nearly as much power as M_2 (1.000) to detect model misfit. A parallel statement, however, cannot be made in the case of the MRS model; C_2 power (.521) actually remained low as compared to M_2 (1.000). In fact, power for C_2 was generally higher given Model 2 (ERS) as compared to Model 3 (MRS), yet this pattern did not hold for M_2.

[Table 8 near here]

[Table 9 near here]

In addition, we also evaluated power under an incorrectly specified response style dimension. That is, we examined power when Model 3 was fit to Model 2 data (Table 10) and vice versa (Table 11). Once again, M_2^* had very low power across all conditions; its highest rejection rate was .068 (when substantive slopes were weak, ERS slopes were strong, and N = 500). Moreover, M_2 behavior closely paralleled its performance under the previous misspecification. The general pattern for M_2 was high power that increased with response style slope strength as well as sample size, and decreased as substantive slopes grew stronger. Also, just as in the previous misspecification, the lowest rejection rate for M_2 (.309) emerged when substantive slopes were strong, response style slopes were weak, and N = 500. C_2, in contrast, performed arguably worse in detecting an incorrectly specified style factor. The highest rejection rate for C_2 was a meager .183 (when substantive and ERS slopes were strong and N = 500), and

C_2 never exhibited a rejection rate higher than one-third the size of the corresponding M_2 rate. Furthermore, while in the earlier power analysis, C_2 power always increased with an increase in sample size, this was not always the case in the currently examined conditions. For example, when Model 3 was fit to Model 2 (Table 10) and substantive and style dimensions were both strong, power dropped from .183 to .137 as sample size increased from N = 500 to N = 1000.

[Table 10 near here]

[Table 11 near here]

We note that interpretations regarding RMSEA in response style MIRT models should be made with caution, as guiding principles for RMSEA in MIRT in general are few and may not be akin to those followed in other contexts. Because RMSEA indices depend on the specific test statistic of interest, fit assessment based on different test statistics may be governed by different criteria, thereby making interpretation not so straightforward. To our knowledge, the literature offers minimal guidance on this matter for M_2 when data are polytomous, and almost none for M_2^* and C_2. For M_2, Maydeu-Olivares and Joe (2014) show that RMSEA behavior is primarily affected by the number of categories, so that as the number of categories increases, the RMSEA value decreases. It is suggested, therefore, that a cutoff criterion for excellent fit can be adapted from the criterion for binary data (.05), such that the relevant criterion becomes $.05/(K-1)$, where K is the number of categories. No similar cutoff criteria have been offered in this context for M_2^* and C_2.

These challenges comprise a substantial barrier when considering the utility of RMSEA as a tool for evaluating the comparative performance of M_2, M_2^*, and C_2 in this study. We nevertheless include, for reference, mean RMSEA values in Tables 8 through 11 to assess per degree of freedom error of approximation (e.g., Cai & Monroe, 2014). If in the case of M_2, one

were to follow the rule set forth by Maydeu-Olivares and Joe (2014), only when response style slopes were weak did mean RMSEA values incorrectly suggest excellent fit (< .0125). Results for M_2^* and C_2 cannot be measured against a cutoff criterion, but are comparable to those found in previous simulation research (Cai & Monroe, 2014). Even given misspecification under which a response style factor is excluded, mean RMSEA values remained low. For M_2^* and C_2, mean RMSEA values tended to decrease slightly at larger sample sizes, suggesting convergence to population RMSEA values that are somewhat smaller.

6. Discussion

In this simulation study, we evaluated the relative performance of limited-information test statistics M_2, M_2^*, and C_2 in the context of the MNRM and response styles. While M_2^* and C_2 showed good calibration under the null across most conditions, M_2 exhibited inflated Type I error rates when substantive slopes were strong. M_2 always had higher power than the other test statistics; its power often approached one except when response style slopes were weak and sample size was smaller. Power to detect model misspecification was close to zero for M_2^* given all conditions and misspecifications. As for C_2, when misspecification under a larger sample size and strong substantive slopes concerned an omitted response style factor with strong slopes, power levels were close to those of M_2; otherwise, C_2 had lower power.

To integrate these findings with previous simulations, while the lack of calibration of M_2 under null conditions is known, the gains in power of M_2 over other limited information test statistics represents a novel finding. Indeed, previous research has often found that C_2 and M_2^* have higher power than M_2 to detect omitted residual dependencies or specific factors of a bifactor model (Cai & Hansen, 2013; Cai & Monroe, 2014). Similarly to Cai and Monroe (2014), we also found that C_2 had greater power than M_2^*. Our findings on the higher power of M_2 are

most likely attributable to a loss of information that occurs when categories are assumed ordinal and are collapsed in univariate and/or bivariate contingency tables under C_2 and M_2^*. Apparently C_2 retains enough information in univariate tables (only bivariate tables are collapsed) to still detect omitted and sometimes misspecified response style factors, yet such information is completely lost under M_2^*.

In making recommendations for practical use, it is important to note that limited information test statistics are overall tests of fit and may not necessarily allow us to pinpoint the source of model misfit. However, given that the test statistics appear sensitive to different kinds of misspecifications, it may be prudent to recommend computation and interpretation of at least M_2 and C_2. We reflect on this issue in the context of the initially presented empirical example (Table 1). In our simulations, M_2^* lacked power and may run out of degrees of freedom with short tests (Cai & Monroe, 2014). Although M_2^* may still retain some power to detect misspecified substantive dimensions (e.g., Cai & Hansen, 2013), C_2 may best it in power for these kinds of misspecifications (Cai & Monroe, 2014). It is thus difficult to make sense of the pattern of results for M_2^* in the empirical example (i.e., rejection of the unidimensional model, but not the MRS model), and it may be more worthwhile to instead interpret M_2 and C_2. Given that C_2 is more powerful to detect misspecifications with substantive dimensions (Cai & Monroe, 2014), but M_2 is more sensitive to response style misspecification (this manuscript), it may be informative when only one test statistic rejects the model as this may be suggesting a particular kind of misspecification. If both reject the model, as is the case with the empirical example, it is more difficult to determine the source of misspecification as it could be misspecification of either style and/or substantive dimensions.

Two caveats to preference for reporting both M_2 and C_2 are noted. Although M_2 experienced some inflated Type I error in our simulations, such rates improve with an increase in sample size and may eventually reach nominal levels, even with strong substantive slopes. In addition, M_2 is known to be slower to compute than C_2, especially when there are many items and categories. Thus, M_2 may be computationally infeasible for a long test, and C_2 may then be the most appropriate fallback option despite its limitations.

Above all, substantive theory should also be considered if no model fits the data and the researcher wishes to engage in some model modification to find a better fitting model. In the original ACME paper (Vachon & Lynam, 2016), all substantive constructs were modeled simultaneously and as independent clusters using a limited information estimation approach, and additional factors representing positively and negatively worded items were added as method factors. Thus, other theoretically plausible models could be considered. If the research question concerns more direct comparisons of alternative response style models, overall fit still provides valuable information, though tests of non-nested models for MIRT models are recently emerging and may be further studied for their use with the MNRM and other polytomous MIRT models (Freeman, 2016; Schneider, Chalmers, Debelak, & Merkle, 2019). Still the researcher may wish to know whether other fit indices such as RMSE and TLI are providing any information about the size of misfit, despite rejection of the models by M_2 and C_2.

In reflecting on RMSEA, little guidance is available regarding interpretation in the case of limited information test statistics with polytomous data, whether ordinal or nominal. We note that this fit index is based on the test statistics themselves and may depend substantially on the number of response categories per item. Our research adds to the literature suggesting that these indices may require different interpretations across M_2-, M_2^*-, and C_2-based model fit evaluation

(Maydeu-Olivares & Joe, 2014; Monroe & Cai, 2015). While a cutoff for close fit has been suggested for RMSEA based on M_2, no absolute cutoff criterion has emerged for M_2^* and C_2, as patterns of performance have been found to be inconsistent across varied numbers of items and categories. As such, the RMSEA values we present were included primarily for reference. While one might be tempted to say that for M_2^* and C_2 the studied models represented minor misspecifications, it may be that RMSEA values for these test statistics should be evaluated with more stringency and the adjusted M_2-based cutoff of .0125 may be too large to use as a rule of thumb. Additional studies are needed before RMSEA values across M_2, M_2^*, and C_2 can be adequately compared, and before better recommendations can be made.

Additional work on other fit indices could be used to supplement limited information test statistics. We did not investigate the performance of TLI as computed by flexMIRT® as this would require additional computations (i.e., estimation of a null model). TLI falls under a class of fit indices (which includes CFI), borrowed from the broader structural equation modeling literature, where the fitted model is implicitly compared to a more restricted null model (or other, usually poorly fitting model; for a review, see Bentler, 2008). In addition, residual-based fit indices (under which standardized root mean square residuals, or SRMR, falls) could also be further developed. For computation of SRMR, while under the assumption of ordinal latent trait models it may be possible to compute both an observed and model-implied correlation matrix among underling variables (e.g., Maydeu-Olivares, 2015), this approach may not work for items with nominal or unordered categories as it is unclear how observed (polychoric) correlations can be computed without some ordinality assumption for the items' categories. Future research may instead draw inspiration on prior work in which residuals of bivariate contingency tables are used with nominal data in order to provide descriptive measures of model fit (e.g., Bacher, 1995).

While our findings represent new knowledge in the area of MIRT model fit assessment, additional questions and limitations remain. First, trends in performance across different sample sizes may become more readily evident with more than two sample size conditions. Of particular interest may be whether C_2 reaches adequate power at larger sample sizes when the type of response style factor is misspecified. Second, although we may expect the Type I calibration of M_2 to get worse with more items and categories, the impact on power is somewhat unpredictable. Still, we would expect a similar pattern of relative performance across test statistics. Next, we also do not know how these statistics behave when univariate frequencies are low (< .05) and when categories are collapsed prior to estimation in order to accommodate such cases. However, presumably M_2 and C_2 will be equally vulnerable to such sparseness. In addition, we only evaluated test statistic performance for models of at most two dimensions. This study design choice was based mainly on computational considerations, and we await advances in accurate computation of limited information test statistics for higher dimensional models. The correlation between response style and substantive factors was zero in the data generating models. However, we have no reason to believe that this limits the generalizability of the findings.[6] Finally, the full range of possible misspecifications that can be reasonably detected by limited information test statistics is largely unknown. Thus, while we focus primarily on misspecification of style and substantive dimensions in our recommendations, we do not know whether limited information test statistics can detect various forms of local dependence (e.g., Liu & Thissen, 2014), omitted

6. In support of this claim, we computed model-implied probabilities for all 5^{10} response patterns for a 10-item (5 categories per item), two-factor (substantive + ERS) model with a .3 correlation among latent dimensions. As an example, fitting alternative misspecified models (a single substantive dimension) to these model-implied proportions typically yielded near zero test statistics for M_2^*, even if a wildly inflated sample size was specified for the fitted model (e.g., close to 1 million). This strategy was deemed not appropriate for the 12-item measures in our empirical example and simulation study, as doing computations on 5^{12} patterns was too RAM and processor intensive.

cross-loadings (e.g., Falk & Monroe, 2018), nonnormal latent trait distributions (Li & Cai, 2018), incorrectly ordered categories (Preston, Reise, Cai, & Hays, 2011), and so on.

It remains to be seen whether these results generalize to other types of scoring function values (Falk & Cai, 2016), the fully unconstrained nominal model, other response style MIRT models (e.g., Bockenholt, 2012, 2017; Thissen-Roe & Thissen, 2013), or more generally to other non-ordinal latent trait models. For example, it is possible that for some items, certain adjacent categories on the substantive dimension are either out of order or effectively indistinguishable by participants, which could also be modeled with the MNRM (Preston et al., 2011). Based on the present results, we would expect that M_2 would have higher power than C_2 and M_2^* to detect such misspecifications. We have no theoretical reason to expect a different pattern of results as all such models violate ordinality as defined by Tutz (2019). That is, M_2^* is expected to lose power to detect misspecification under such models, and we would generally expect a similar relative performance between M_2 and C_2.

7. References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 466-485. https://doi.org/10.1111/bmsp.12169

Agresti, A. (1996). An Introduction to Categorical Data Analysis. New York, NY: Wiley. https://doi.org/10.1002/9780470114759

Bacher, J. (1995). Goodness-of-fit measures for multiple correspondence analysis. *Quality & Quantity, 29,* 1-16. https://doi.org/10.1007/BF01107980

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research, 38,* 143-156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bentler, P. M. (2008). *EQS structural equation modeling software.* Encino, CA: Multivariate Software.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51. https://doi.org/10.1007/BF02291411

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika, 46,* 443-459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*(2), 179-197. https://doi.org/10.1007/BF02291262

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods, 17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods, 22*(1), 69–83. https://doi.org/10.1037/met0000106

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement, 33*, 335-352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods, 19*(4), 528–541. https://doi.org/10.1037/met0000016

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*, 814–833. https://doi.org/10.1177/0013164410388411

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

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245-276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data (*CRESST *Report 839)*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

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

Dowling, N. M., Bolt, D. M., Deng, S., & Li, C. (2016). Measurement and control of bias in patient reported outcomes using multidimensional item response theory. *BMC Medical Research Methodology, 16,* 63. https://doi.org/10.1186/s12874-016-0161-z

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21,* 328-347. https://doi.org/10.1037/met0000059

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Quantitative Psychology and Measurement, 11 (72),* 1-17. https://doi.org/10.3389/fpsyg.2020.00072

Falk, C. F., & Monroe, S. (2018). On Lagrange multiplier tests in multidimensional item response theory: Information matrices and model misspecification. *Educational and Psychological Measurement, 78*(4), 653-678. https://doi.org/10.1177/0013164417714506

Freeman, L. (2016). *Assessing model-data fit for compensatory and non-compensatory multidimensional item response models using Vuong and Clarke statistics.* Unpublished doctoral dissertation, University of Wisconsin – Milwaukee.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42. https://doi.org/10.1037/1040-3590.4.1.26

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75,* 393-419. https://doi.org/10.1007/s11336-010-9165-5

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics, 35*(1), 92-114. https://doi.org/10.3102/1076998609340529

Jonas, K. G., & Markon, K. E. (2018). Modeling response style using vignettes and person-specific item response theory. *Applied Psychological Measurement, 00*, 1-15. https://doi.org/10.1177/0146621618798663

Ju, U., & Falk, C. F. (2019). Modeling response styles in cross-country self-reports: An application of a multilevel multidimensional nominal response model. *Journal of Educational Measurement, 56,* 169-191. https://doi.org/10.1111/jedm.12205

Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multiscale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, *49*(2), 161-177. https://doi.org/10.1080/00273171.2013.866536

Kieruj, N. D., & Moors, G. (2013). Response style behavior: question format dependent or personal style?. *Quality & Quantity*, *47*(1), 193-211. https://doi.org/10.1007/s11135-011-9511-4

Li, Z., & Cai, L. (2018). Summed score likelihood–based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, *78*(5), 857-886. https://doi.org/10.1177/0013164417717024

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. *British Journal of Mathematical and Statistical Psychology*, *67*(3), 496-513. https://doi.org/10.1111/bmsp.12030

Liu, C. W., & Wang, W. C. (2019). A general unfolding IRT model for multiple response styles. *Applied Psychological Measurement*, *43*(3), 195-210. https://doi.org/10.1177/0146621618762743

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge. https://doi.org/10.4324/9780203501207

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. https://doi.org/10.1007/BF02296272

Maurer, T. J., Mitchell, D. R. D., & Barbeite, F. G. (2002). Predictors of attitudes towards a 360-degree feedback system and involvement in post-feedback management development activity. *Journal of Occupational and Organizational Psychology, 75*, 87-107. https://doi.org/10.1348/096317902167667

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71-101. https://doi.org/10.1080/15366367.2013.831680

Maydeu-Olivares, A. (2015). Evaluating fit in IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111-127). New York: Routledge.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables. *Journal of the American Statistical Association, 100*, 1009-1020. https://doi.org/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713–732. https://doi.org/10.1007/s11336-005-1295-9

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305–328. https://doi.org/10.1080/00273171.2014.911075

Monroe, S., & Cai, L. (2015). Evaluating structural equation models for categorical outcomes: a new test statistic and a practical challenge of interpretation. *Multivariate Behavioral Research, 50*, 569-583. https://doi.org/10.1080/00273171.2015.1032398

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176. https://doi.org/10.1177/014662169201600206

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightman (Eds.), *Measurement of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press. https://doi.org/10.1016/b978-0-12-590241-0.50006-x

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224-239). New York, NY: Guilford Press.

Preston, K., Reise, S., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate category discrimination in the PROMIS emotional distress item pools. *Educational and Psychological Measurement, 71*, 523-550. https://doi.org/10.1177/0013164410382250

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 17*. https://doi.org/10.1007/BF03372160

Schneider, L., Chalmers, R. P., Debelak, R., & Merkle, E. C. (2019). Model selection of nested and non-nested item response models using Vuong tests. *Multivariate Behavioral Research, 54,* 1-21. https://doi.org/10.1080/00273171.2019.1664280

Schneider, S. (2018). Extracting response style bias from measures of positive and negative affect in aging research. *The Journals of Gerontology: Series B, 73*(1), 64-74. https://doi.org/10.1093/geronb/gbw103

Stone, A. A., Schneider, S., Junghaenel, D. U., & Broderick, J. E. (2019). Response styles confound the age gradient of four health and well-being outcomes. *Social Science Research, 78,* 215-225. https://doi.org/10.1016/j.ssresearch.2018.12.004

Tafarodi, R. W., & Swann, W. B. (2001). Two-dimensional self-esteem: theory and measurement. *Personality and Individual Differences, 31*, 653-673. https://doi.org/10.1016/S0191-8869(00)00169-0

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume one: Models* (pp. 51–73). Boca Raton, FL: Chapman & Hall/CRC.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Developments and applications* (pp. 43-75). New York, NY: Taylor & Francis.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics, 38*(5), 522–547. https://doi.org/10.3102/1076998613481500

Tutz, G. (2019). What is an Ordinal Latent Trait Model? *arXiv preprint arXiv:1902.06303*.

Vachon, D. D., & Lynam, D. R. (2016). Fixing the problem with empathy: Development and validation of the Affective and Cognitive Measure of Empathy scale. *Assessment, 23*, 135-149. https://doi.org/10.1177/1073191114567941

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*(5), 352-364. https://doi.org/10.1027/1015-5759/a000291

Table 1. Model Fit Results for ACME Affective Resonance (AR) Data.

Model	Statistic	*df*	Value	*p*	RMSEA	TLI	AIC
AR	M_2	1044	7656	0.0001	0.09	0.68	20445
AR+ERS	M_2	1031	4572	0.0001	0.07	0.83	19564
AR+MRS	M_2	1031	3742	0.0001	0.06	0.87	19362
AR+ERS+MRS	M_2	1017	2125	0.0001	0.04	0.95	18617
AR	C_2	54	689	0.0001	0.12	0.91	20445
AR+ERS	C_2	41	196	0.0001	0.07	0.97	19564
AR+MRS	C_2	41	337	0.0001	0.09	0.95	19362
AR+ERS+MRS	C_2	27	97	0.0001	0.06	0.98	18617
AR	M_2^*	18	152	0.0001	0.10	0.97	20445
AR+ERS	M_2^*	5	9	0.0920	0.03	1.00	19564
AR+MRS	M_2^*	5	16	0.0058	0.05	0.99	19362
AR+ERS+MRS	M_2^*	neg					

Table 2. Bivariate Observed (and Expected) Marginal Proportions for Items 5 and 6 of the Affective Resonance Subscale

		Item 6 Response Category					Marginal Proportion for Item 5
		0	1	2	3	4	
Item 5 Response Category	0	.315 (.290)	.085 (.117)	.021 (.030)	.007 (.004)	.001 (.001)	.430 (.442)
	1	.110 (.133)	.168 (.117)	.030 (.057)	.017 (.013)	0 (.003)	.325 (.322)
	2	.019 (.032)	.034 (.052)	.097 (.043)	.015 (.016)	0 (.005)	.164 (.148)
	3	.005 (.006)	.020 (.015)	.020 (.020)	.012 (.012)	.005 (.006)	.062 (.060)
	4	.001 (.001)	.002 (.004)	.005 (.008)	.001 (.008)	.009 (.007)	.019 (.027)
Marginal Proportion for Item 6		.450 (.461)	.309 (.305)	.173 (.158)	.054 (.053)	.015 (.022)	

Table 3. Reduced Univariate and Bivariate Sample Proportions (Model-Implied Probabilities) for Items 5 and 6 of Affective Resonance Example

	Item 6	Mean for Item 5
Item 5	1.38 (1.33)	.915 (.908)
Mean for Item 6	.875 (.869)	

Note: The reduced bivariate statistic is an item-pair moment.

Table 4. Summary of Available Information for Limited Information Test Statistics

	Statistic		
	M_2	M_2^*	C_2
1st Order Collapsed?	No	Yes	No
2nd Order Collapsed?	No	Yes	Yes
1st Order Info	$\sum_{j=1}^{n}(K_j - 1)$	n	$\sum_{j=1}^{n}(K_j - 1)$
2nd Order Info	$\sum_{j=1}^{n-1}\sum_{h=j+1}^{n}(K_j - 1)(K_h - 1)$	$n(n-1)/2$	$n(n-1)/2$
1st Order Info, Equal K	$n(K-1)$	n	$n(K-1)$
2nd Order Info, Equal K	$(n(n-1)/2)(K-1)^2$	$n(n-1)/2$	$n(n-1)/2$
flexMIRT® keyword	M_2 = "Full"	M_2^* = "Ordinal"	C_2 = "Mixed"

Table 5. Parameters for the Model 2 and Model 3 Generating Model with Strong Substantive Slopes and Strong Response Style Slopes

Item i	Intercept Contrasts				Slopes		Standardized Loadings	
	γ_{i1}	γ_{i2}	γ_{i3}	γ_{i4}	α_{iSU}	α_{iRS}	λ_{iSU}	λ_{iRS}
1	1.00	1.25	0.15	-0.10	1.00	1.20	0.43	0.52
2	-1.00	1.00	0.20	-0.10	1.00	1.20	0.43	0.52
3	0.00	1.50	0.25	0.10	1.00	1.20	0.43	0.52
4	1.00	1.25	0.15	-0.10	1.00	1.35	0.42	0.56
5	-1.00	1.00	0.20	-0.10	1.00	1.35	0.42	0.56
6	0.00	1.50	0.25	0.10	1.00	1.35	0.42	0.56
7	1.00	1.25	0.15	-0.10	1.20	1.30	0.49	0.53
8	-1.00	1.00	0.20	-0.10	1.20	1.30	0.49	0.53
9	0.00	1.50	0.25	0.10	1.20	1.30	0.49	0.53
10	1.00	1.25	0.15	-0.10	1.20	1.40	0.48	0.56
11	-1.00	1.00	0.20	-0.10	1.20	1.40	0.48	0.56
12	0.00	1.50	0.25	0.10	1.20	1.40	0.48	0.56

Table 6. Simulation Results: Null Conditions, Model 1

Slopes	*N*	Statistic	*df*	Mean	Variance	α = .01	α = .05	α = .10	RMSEA
weak	500	M_2	1044	1046.895	2090.264	0.013	0.058	0.111	0.004
		M_2^*	18	17.987	38.557	0.013	0.054	0.106	0.010
		C_2	54	53.989	105.860	0.012	0.051	0.094	0.008
weak	1000	M_2	1044	1042.556	2042.875	0.010	0.036	0.094	0.003
		M_2^*	18	18.330	39.924	0.014	0.059	0.108	0.008
		C_2	54	54.243	118.535	0.017	0.060	0.109	0.006
strong	500	M_2	1044	1044.372	3752.639	0.052	0.103	0.134	0.004
		M_2^*	18	17.916	35.583	0.012	0.047	0.096	0.010
		C_2	54	53.700	105.105	0.010	0.050	0.092	0.008
strong	1000	M_2	1044	1044.074	3176.634	0.032	0.087	0.151	0.003
		M_2^*	18	18.012	36.706	0.008	0.053	0.106	0.007
		C_2	54	53.860	115.562	0.011	0.046	0.105	0.005

Table 7. Simulation Results: Null Conditions, Model 2

Slopes	*N*	Statistic	*df*	Mean	Variance	α = .01	α = .05	α = .10	RMSEA
weak, weak	500	M_2	1032	1033.673	2055.547	0.013	0.058	0.105	0.004
		M_2^*	6	5.914	11.484	0.007	0.042	0.090	0.012
		C_2	42	41.949	79.910	0.011	0.035	0.088	0.008
weak, weak	1000	M_2	1032	1033.069	2095.982	0.007	0.055	0.103	0.003
		M_2^*	6	5.968	11.660	0.007	0.050	0.097	0.009
		C_2	42	42.340	89.237	0.012	0.063	0.111	0.006
weak, strong	500	M_2	1032	1035.963	1964.880	0.016	0.058	0.119	0.004
		M_2^*	6	5.869	10.292	0.004	0.041	0.087	0.012
		C_2	42	41.978	89.878	0.010	0.055	0.099	0.008
weak, strong	1000	M_2	1032	1033.992	2090.545	0.012	0.052	0.113	0.003
		M_2^*	6	5.956	11.499	0.007	0.050	0.096	0.008
		C_2	42	42.282	84.212	0.005	0.053	0.106	0.006
strong, weak	500	M_2	1032	1030.175	3644.616	0.046	0.108	0.151	0.004
		M_2^*	6	5.942	11.454	0.006	0.049	0.089	0.012
		C_2	42	41.575	80.007	0.010	0.043	0.084	0.008
strong, weak	1000	M_2	1032	1032.383	2604.663	0.020	0.079	0.140	0.003
		M_2^*	6	5.998	13.236	0.010	0.062	0.101	0.009
		C_2	42	42.253	86.293	0.009	0.056	0.108	0.006
strong, strong	500	M_2	1032	1031.336	3540.547	0.044	0.100	0.165	0.004
		M_2^*	6	5.870	11.549	0.009	0.041	0.092	0.012
		C_2	42	41.561	83.146	0.015	0.044	0.083	0.008
strong, strong	1000	M_2	1032	1031.753	2540.759	0.018	0.068	0.126	0.003
		M_2^*	6	5.928	10.879	0.004	0.046	0.103	0.008
		C_2	42	42.494	84.620	0.006	0.052	0.097	0.006

Note. Strength of slopes for the substantive dimension is listed first in each condition, followed by strength of slopes for the response style dimension.

Table 8. Simulation Results: Power, Model 1 Fitted to Model 2 Data

Slopes	*N*	Statistic	*df*	α = .01	α = .05	α = .10	RMSEA
weak, weak	500	M_2	1044	0.221	0.418	0.545	0.010
		M_2^*	18	0.012	0.049	0.098	0.010
		C_2	54	0.022	0.098	0.171	0.011
weak, weak	1000	M_2	1044	0.655	0.830	0.896	0.010
		M_2^*	18	0.007	0.065	0.108	0.008
		C_2	54	0.035	0.112	0.193	0.008
weak, strong	500	M_2	1044	1.000	1.000	1.000	0.045
		M_2^*	18	0.017	0.068	0.137	0.012
		C_2	54	0.491	0.699	0.791	0.030
weak, strong	1000	M_2	1044	1.000	1.000	1.000	0.046
		M_2^*	18	0.020	0.085	0.141	0.009
		C_2	54	0.862	0.945	0.975	0.029
strong, weak	500	M_2	1044	0.188	0.365	0.472	0.010
		M_2^*	18	0.006	0.031	0.071	0.009
		C_2	54	0.045	0.130	0.207	0.012
strong, weak	1000	M_2	1044	0.648	0.824	0.893	0.010
		M_2^*	18	0.009	0.046	0.088	0.006
		C_2	54	0.077	0.213	0.335	0.012
strong, strong	500	M_2	1044	1.000	1.000	1.000	0.046
		M_2^*	18	0.004	0.020	0.045	0.006
		C_2	54	0.960	0.986	0.994	0.049
strong, strong	1000	M_2	1044	1.000	1.000	1.000	0.048
		M_2^*	18	0.001	0.017	0.046	0.005
		C_2	54	1.000	1.000	1.000	0.049

Note. Strength of slopes for the substantive dimension is listed first in each condition, followed by strength of slopes for the response style dimension.

Table 9. Simulation Results: Power, Model 1 Fitted to Model 3 Data

Slopes	*N*	Statistic	*df*	α = .01	α = .05	α = .10	RMSEA
weak, weak	500	M_2	1044	0.361	0.601	0.714	0.012
		M_2^*	18	0.014	0.068	0.126	0.011
		C_2	54	0.014	0.052	0.122	0.008
weak, weak	1000	M_2	1044	0.873	0.964	0.975	0.012
		M_2^*	18	0.020	0.053	0.111	0.007
		C_2	54	0.020	0.078	0.140	0.006
weak, strong	500	M_2	1044	1.000	1.000	1.000	0.052
		M_2^*	18	0.027	0.080	0.141	0.012
		C_2	54	0.113	0.307	0.425	0.018
weak, strong	1000	M_2	1044	1.000	1.000	1.000	0.051
		M_2^*	18	0.018	0.069	0.129	0.008
		C_2	54	0.297	0.521	0.641	0.018
strong, weak	500	M_2	1044	0.176	0.319	0.437	0.009
		M_2^*	18	0.010	0.051	0.095	0.009
		C_2	54	0.011	0.061	0.107	0.008
strong, weak	1000	M_2	1044	0.526	0.753	0.829	0.010
		M_2^*	18	0.007	0.042	0.084	0.007
		C_2	54	0.011	0.064	0.128	0.006
strong, strong	500	M_2	1044	1.000	1.000	1.000	0.046
		M_2^*	18	0.005	0.032	0.065	0.008
		C_2	54	0.373	0.598	0.724	0.027
strong, strong	1000	M_2	1044	1.000	1.000	1.000	0.046
		M_2^*	18	0.004	0.036	0.070	0.006
		C_2	54	0.819	0.917	0.953	0.028

Note. Strength of slopes for the substantive dimension is listed first in each condition, followed by strength of slopes for the response style dimension.

Table 10. Simulation Results: Power, Model 3 Fitted to Model 2

Slopes	*N*	Statistic	*df*	α = .01	α = .05	α = .10	RMSEA
weak, weak	500	M_2	1032	0.185	0.382	0.497	0.010
		M_2^*	6	0.013	0.056	0.100	0.012
		C_2	42	0.011	0.067	0.147	0.010
weak, weak	1000	M_2	1032	0.619	0.795	0.877	0.010
		M_2^*	6	0.009	0.052	0.107	0.009
		C_2	42	0.025	0.085	0.146	0.007
weak, strong	500	M_2	1032	1.000	1.000	1.000	0.045
		M_2^*	6	0.015	0.068	0.131	0.014
		C_2	42	0.054	0.168	0.267	0.014
weak, strong	1000	M_2	1032	1.000	1.000	1.000	0.044
		M_2^*	6	0.019	0.062	0.112	0.010
		C_2	42	0.036	0.129	0.223	0.010
strong, weak	500	M_2	1032	0.174	0.333	0.442	0.009
		M_2^*	6	0.009	0.042	0.100	0.012
		C_2	42	0.024	0.086	0.149	0.011
strong, weak	1000	M_2	1032	0.605	0.793	0.872	0.010
		M_2^*	6	0.009	0.045	0.094	0.009
		C_2	42	0.042	0.135	0.224	0.009
strong, strong	500	M_2	1032	1.000	1.000	1.000	0.045
		M_2^*	6	0.005	0.028	0.043	0.008
		C_2	42	0.106	0.183	0.265	0.015
strong, strong	1000	M_2	1032	1.000	1.000	1.000	0.046
		M_2^*	6	0.003	0.023	0.062	0.006
		C_2	42	0.052	0.137	0.205	0.008

Note. Strength of slopes for the substantive dimension is listed first in each condition, followed by strength of slopes for the response style dimension.

Table 11. Simulation Results: Power, Model 2 Fitted to Model 3

Slopes	*N*	Statistic	*df*	α = .01	α = .05	α = .10	RMSEA
weak, weak	500	M_2	1032	0.326	0.561	0.670	0.012
		M_2^*	6	0.013	0.040	0.089	0.012
		C_2	42	0.009	0.054	0.120	0.009
weak, weak	1000	M_2	1032	0.845	0.950	0.974	0.011
		M_2^*	6	0.015	0.060	0.096	0.009
		C_2	42	0.015	0.066	0.119	0.006
weak, strong	500	M_2	1032	1.000	1.000	1.000	0.051
		M_2^*	6	0.015	0.058	0.112	0.013
		C_2	42	0.021	0.102	0.164	0.011
weak, strong	1000	M_2	1032	1.000	1.000	1.000	0.050
		M_2^*	6	0.009	0.047	0.099	0.009
		C_2	42	0.017	0.084	0.149	0.007
strong, weak	500	M_2	1032	0.184	0.309	0.420	0.009
		M_2^*	6	0.012	0.048	0.086	0.012
		C_2	42	0.011	0.041	0.102	0.008
strong, weak	1000	M_2	1032	0.499	0.729	0.814	0.010
		M_2^*	6	0.009	0.046	0.095	0.009
		C_2	42	0.011	0.050	0.100	0.006
strong, strong	500	M_2	1032	1.000	1.000	1.000	0.044
		M_2^*	6	0.012	0.035	0.084	0.012
		C_2	42	0.011	0.045	0.086	0.008
strong, strong	1000	M_2	1032	1.000	1.000	1.000	0.044
		M_2^*	6	0.003	0.034	0.077	0.008
		C_2	42	0.007	0.052	0.095	0.005

Note. Strength of slopes for the substantive dimension is listed first in each condition, followed by strength of slopes for the response style dimension.