Title:
A flexible full-information approach to the modeling of response styles

Authors:
Carl F. Falk
Li Cai

Journal publication date:
2016

Published in:
Psychological Methods, 21(3), 328-347

IES grant information:
Grant number R305D140046
Funded by National Center for Education Research (NCER)

A flexible full-information approach to the modeling of response styles

Carl F. Falk
Michigan State University

Li Cai
University of California, Los Angeles

Accepted at *Psychological Methods*

This research was partially supported by a Social Sciences and Humanities Research Council of Canada Postdoctoral Fellowship awarded to Carl F. Falk. Li Cai's research is partially supported by a grant from the Institute of Education Sciences (R305D140046).

Address all correspondence to: Carl F. Falk, Measurement and Quantitative Methods, Michigan State University, 458 Erickson Hall, East Lansing, MI, USA 48824.
Email: cffalk@gmail.com. Phone: 562.221.7538.

A flexible full-information approach to the modeling of response styles

Abstract

In this paper, we present a flexible full-information approach to modeling multiple user-defined response styles across multiple constructs of interest. The model is based on a novel parameterization of the multidimensional nominal response model that separates estimation of overall item slopes from the scoring functions (indicating the order of categories) for each item and latent trait. This feature allows for the definition of response styles to vary across items as well as overall item slopes that vary across items for both substantive and response style dimensions. The model is compared with similar approaches using examples from the smoking initiative of NIH's Patient Reported Outcomes Measurement Information System. A small set of simulations show that the estimation approach is able to recover model parameters, factor scores, and reasonable estimates of standard errors. Furthermore, these simulations suggest that failing to include response style factors (when present in the data generating model) has adverse consequences for substantive trait factor score recovery.

Keywords: Response styles; multidimensional item response theory; nominal response model

1 Introduction

Researchers often ask their participants to complete survey measures that include Likert-type items (e.g., rating one's agreement to a statement on a scale from *1 - Strongly Disagree* to *7 - Strongly Agree*). The use of such items is pervasive throughout the social sciences, including social and personality psychology (Paulhus & Vazire, 2007), clinical psychology (Morey, 1991), patient reported outcomes (Hansen et al., 2014), and so on. However, individuals are sometimes thought to exhibit different response styles; that is, they tend to use the response options in different ways (Baumgartner & Steenkamp, 2001; Paulhus, 1991). Common examples include extreme response style (ERS; tendency to use the endpoints, e.g., 1 or 7), midpoint responding (MRS; tendency to use of the middle response category; e.g., 4) and acquiescence (ARS; a tendency to agree with items; Baumgartner & Steenkamp, 2001). Additional response styles may have an underlying motivational component, such as responding to all items in a socially desirable way (SDR; Kuncel & Tellegen, 2009; Paulhus, 1991) to make oneself look good to others.

Although there are many different approaches to modeling or measuring response styles (Chen, Lee, & Stevenson, 1995; Cheung & Rensvold, 2000; Clarke, 2001; Greenleaf, 1992; Fischer, 2004), recent advances include the development of full-information methods that use the same items to simultaneously assess substantive constructs and response styles. For instance, responses to such items may involve multiple decision processes that depend on the latent variables for the construct of interest and response styles (e.g., ERS and MRS), and can be formally modeled using multidimensional item response theory (e.g., Böckenholt, 2012, 2014; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013). Wang and colleagues (Jin & Wang, 2014; Wang, Wilson, & Shih, 2006; Wang & Wu, 2011) have presented extensions of the rating scale, partial credit, and generalized partial credit models that can include either random thresholds across persons or a multiplicative person parameter representing ERS that affects item thresholds. Other approaches allow for heterogeneous thresholds across

persons to account for ERS using a graded response model (Johnson, 2003; Rossi, Gilula, & Allenby, 2001), ERS using mixed Rasch models (e.g., Wetzel, Carstensen, & Böhnke, 2013), or ERS and ARS using an unfolding model (Javaras & Ripley, 2007). Using a multidimensional nominal response model (MNRM), response style factors have been added in a compensatory manner to affect the probability of choosing certain categories; such examples include discrete latent traits (or latent classes) representing ERS (Kieruj & Moors, 2010; Moors, 2003, 2004; Morren, Gelissen, & Vermunt, 2011, 2012), ERS and ARS (Kieruj & Moors, 2013), response style classes (van Rosmalen, van Herk, & Groenen, 2010), or continuous latent traits for ERS (Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010).

While these approaches are all important developments, a review of the literature reveals some limitations in the flexibility of the types of response styles that can be accommodated, and the capacity to simultaneously model multiple response styles and multiple substantive constructs. For instance, with the exception of a few approaches (Böckenholt, 2012; Bolt & Newton, 2011; van Rosmalen et al., 2010), most of the above developments are specialized for modeling only one or two response styles at a time - usually ERS, MRS, and/or ARS. In some cases, ERS and MRS are considered opposite poles of the same underlying dimension and cannot be separated (e.g., Jin & Wang, 2014; Thissen-Roe & Thissen, 2013). In some cases where two response styles are modeled, a simplifying assumption is often made that the response style dimension affects all items equally (i.e., loadings are equal across items; Böckenholt, 2012; Bolt & Newton, 2011; Plieninger & Meiser, 2014; Kieruj & Moors, 2013, for an exception see Khorramdel & von Davier, 2014). Even if this assumption were correct, these applications exclude the possibility where a researcher would like to explicitly test this assumption. While decision process models may offer flexibility in the types of response styles modeled, they have so far been restricted to either models with equal loadings across items (Böckenholt, 2012; Plieninger & Meiser, 2014) or demonstrations modeling only ERS and MRS (e.g.,

Khorramdel & von Davier, 2014; Thissen-Roe & Thissen, 2013).

Computational reasons have also limited the number of constructs modeled. Response style models have been demonstrated with Maximum Likelihood estimation via the EM algorithm coupled with numerical integration using either Gaussian or Monte Carlo methods, or the use of Bayesian MCMC methods. These approaches can be computationally prohibitive (see Cai, 2010a). As a result, the most number of correlated continuous latent traits that we have seen in the above research is only three (Bolt & Newton, 2011; Kieruj & Moors, 2013). To achieve the modeling of additional traits, some research has substituted latent classes with a discrete number of ordered levels instead of continuous latent traits (Khorramdel & von Davier, 2014; Kieruj & Moors, 2013; Moors, 2003, 2004; van Rosmalen et al., 2010), or estimate substantive traits that are uncorrelated (e.g., Thissen-Roe & Thissen, 2013).

To address these limitations, the present research combines the innovative approach of using continuous latent traits and the MNRM by Bolt and colleagues (Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010) with a novel parameterization of the MNRM (Thissen, Cai, & Bock, 2010; Thissen & Cai, in press) and efficient estimation via the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a, 2010b). The model subsumes work by Bolt and colleagues, and is able to handle multiple user-defined response style factors across multiple substantive traits. In contrast to previous research with the MNRM, our implementation includes overall item slopes for each factor (rather than slopes for individual categories as in the original MNRM), which allows researchers to fit and test models where the effect of response styles may be constant or varying across items. In addition, the model allows users to define a particular response style in a *different* way across items, meaning that the response style can be defined by the choice of different categories across items as is the case with recent formulations of socially desirable responding (Kuncel & Tellegen, 2009). In this paper, we demonstrate the simultaneous modeling of ERS and MRS, and we outline how the model may also

be used for ARS, SDR, and other arbitrarily defined response style factors.

Additionally, MH-RM allows for estimation of a high number of continuous correlated latent traits and can result in concurrent estimation of parameter standard errors. In some cases, this approach may be preferable to the use of latent classes or estimation of only a few latent traits. To the extent that the underlying latent traits may be continuous, use of latent classes with few ordinal levels may be a simplifying approximation. Furthermore, it has been recommended that if response styles represent stable personality traits that affect all items, their proper assessment is best achieved by including a large number of items or substantive constructs (Bolt & Newton, 2011; Greenleaf, 1992).

Our primary goal is to present the MNRM modeling framework and demonstrate its flexibility. To this end, Section 2 presents a motivating example for the development of the model. In Section 3, we present the reparameterized MNRM, examples of response functions under various response styles, and compare our modeling approach with previous uses of the MNRM. In Section 4, our approach is demonstrated on empirical data. In Section 5, we present two small simulation studies that evaluate the performance of our approach. Section 6 concludes with closing remarks and future directions.

2 Empirical Example Data

To motivate development of the model, imagine we are left with the following problem using data from the PROMIS®smoking initiative (Hansen et al., 2014). The PROMIS® cigarette smoking domain includes six correlated constructs (nicotine dependence, hedonic benefits, coping benefits, social benefits, psychosocial risks, and health risks), each measured by 12 to 27 5-point Likert-type items (all positively keyed). The item banks were refined using a combination of exploratory and confirmatory item factor analytic methods from a much larger original pool of items developed and reviewed using standard PROMIS® protocols. Each bank is calibrated on data from 4,201 daily (109 items) and 1,183 non-daily smokers (107 items) collected online by Harris Interactive.

Imagine that we want to perform item calibration across both groups and all six

dimensions simultaneously while probing for the presence of response styles across all items (see Figure 1). For illustration, let us just consider the possibility that ERS *or* MRS *or* both response styles are present in the data. For example, we may hypothesize that some participants have a tendency to use just the endpoints of the scale (ERS) or the midpoint of the scale (MRS) - response strategies that may reduce cognitive burden in making more nuanced choices among the response options (e.g., Kieruj & Moors, 2010). We also question whether ERS and MRS are distinct uncorrelated dimensions or are just opposite poles of the same response style dimension (a tendency to use categories near the middle of the scale versus the endpoints, which we label EMRS). Based on research suggesting that item-level features may affect the prevalence of response styles, we also wonder whether response styles affect all items equally or whether some items are more prone to response styles than others (e.g., perhaps items that require more cognitive effort in order to achieve a nuanced response; Krosnick, 1991).

The need to model MRS separately from ERS precludes approaches that have been developed exclusively for ERS or consider ERS and MRS to only be on opposite poles of the same dimension (e.g., Jin & Wang, 2014; Thissen-Roe & Thissen, 2013). Furthermore, we wish to fit models with eight correlated continuous latent traits and response style loadings that are fixed or vary across items. Both the high dimensionality and varying response style loadings have not been demonstrated in previous uses of the MNRM (Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010). Using this empirical example, in the next section we formalize how the MNRM can be used to accommodate the above modeling situations, as well as other arbitrarily defined response styles.

3 The Proposed Modeling Approach

3.1 The Multidimensional Nominal Response Model (MNRM)

Takane and de Leeuw (1987) presented a multivariate generalization of Bock's nominal response model (NRM; Bock, 1972). In econometrics, McFadden (1974) is generally credited with the development of the multinomial logistic formalism inherent in the

NRM. Johnson and Bolt (2010, see also Moors 2003, 2004) presented the NRM again as useful for response styles.

To introduce notation, consider $i = 1, \ldots, N$ independent subjects who respond to $j = 1, \ldots, n$ items. Let Y_{ij} be a discrete random variable representing observed item response for subject i on item j and y_{ij} is its realization. Let there be $k = 1, \ldots, K_j$ possible ordered response options for item j. Consider also $\mathbf{x}_i$ as a $D \times 1$ vector representing subject i's factor scores on $d = 1, \ldots, D$ latent dimensions. Typically, the latent dimensions are assumed multivariate normal, $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with a certain covariance structure among the latent traits, $\boldsymbol{\Sigma}$. Let $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_N')'$ be an $N \times D$ matrix containing as its rows each individual's vector of factor scores and let $\mathbf{Y}$ be an $N \times n$ matrix of item responses where the (i, j)th entry is Y_{ij}. For now, if we drop item and subject subscripts to avoid notational clutter, and note that item parameters and factor scores may vary across items and subjects, then the category response functions under the original MNRM are based on the multinomial logistic distribution:

$$P(Y = k|\mathbf{x}, \tilde{\mathbf{a}}, \mathbf{c}) = \frac{\exp(\tilde{\mathbf{a}}_k'\mathbf{x} + c_k)}{\sum_{m=1}^{K} \exp(\tilde{\mathbf{a}}_m'\mathbf{x} + c_m)} \tag{1}$$

where $\tilde{\mathbf{a}}_k$ is a $D \times 1$ vector of slopes that represents loadings of category k on the D latent variables and c_k is the intercept for category k. All slopes and intercepts are contained in $\tilde{\mathbf{a}}$ and $\mathbf{c}$, respectively.

Thissen and Cai (in press) present the following reparameterization of the MNRM:

$$P(Y = k|\mathbf{x}, \mathbf{a}, \mathbf{S}, \mathbf{c}) = \frac{\exp([\mathbf{a} \circ \mathbf{s}_k]'\mathbf{x} + c_k)}{\sum_{m=1}^{K} \exp([\mathbf{a} \circ \mathbf{s}_m]'\mathbf{x} + c_m)} \tag{2}$$

where $\mathbf{x}$ and k are defined as before, $\mathbf{a}$ is a vector of slope parameters of length D, and $\mathbf{c}$ is a vector of intercept parameters of length K. The symbol $\circ$ denotes the Schur (or entrywise) product. As before, c_k is the kth element in the vector $\mathbf{c}$, or the intercept that corresponds to category k. $\mathbf{S}$ is a $D \times K$ matrix that contains scoring function values

with $\mathbf{s}_k$ pertaining to the kth column of $\mathbf{S}$. In other words, each column of $\mathbf{S}$ corresponds to a particular category for the item and each row of $\mathbf{S}$ corresponds to a particular latent dimension. This model is similar to one presented by Thissen et al. (2010) and implemented in *IRTPRO* (Cai, Thissen, & du Toit, 2011), *flexMIRT*®(Cai, 2013) and the *mirt* package (Chalmers, 2012) except that the scoring functions in the present approach may vary across latent dimensions. Whereas various submodels of (1) have been used to model response styles (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010), to our knowledge the parameterization in (2) has not and is the focus of the current paper. A module implementing the new model is scheduled for the next major update in *flexMIRT*®sometime late 2015 or early 2016.

3.2 Scoring Functions for the Modeling of Response Styles

The key insight useful for the MNRM in modeling response styles under both of these parameterizations lies in the ability to estimate or fix the order of the categories and interpret how they relate to each latent trait. For the model in (1), the order of categories and the strength of their relation to the latent trait is determined by the category slopes, $\tilde{\mathbf{a}}$. For a 5-category item with ordered responses, slopes for a single dimension proportional to $[\tilde{a}_{d,1} \;\; \tilde{a}_{d,2} \;\; \tilde{a}_{d,3} \;\; \tilde{a}_{d,4} \;\; \tilde{a}_{d,5}] = [0 \;\; 1 \;\; 2 \;\; 3 \;\; 4]$, with $\tilde{a}_{d,k}$ representing the slope for dimension d and category k, would represent graded responses similar to the partial credit (PC) or generalized partial credit (GPC) models (Masters, 1982; Muraki, 1992). In (2), the order of categories is determined by the scoring functions, $\mathbf{S}$. Each row of $\mathbf{S}$ determines the order of the categories for the item and how they relate to that particular latent trait. For example, $[s_{d,1} \;\; s_{d,2} \;\; s_{d,3} \;\; s_{d,4} \;\; s_{d,5}] = [0 \;\; 1 \;\; 2 \;\; 3 \;\; 4]$, with $s_{d,k}$ representing row d and column k of $\mathbf{S}$, is equivalent to the PC or GPC models. The difference in these parameterizations is that under (2) the order of categories (scoring functions), $\mathbf{S}$, are separated from the overall item slopes, $\mathbf{a}$, allowing researchers to fix the order of categories for the dimension according to theoretically interesting or hypothesized values, but estimate overall item slopes for that dimension that may vary across

items. This represents our preferred approach for modeling response styles and allows response styles to have slopes that differ across items. In what follows, we primarily focus on this latter parameterization unless noted otherwise.

To see how response styles may be defined and affect item responses, we consider example response functions for a single 5-category hedonic benefits item ("Smoking makes me feel content"). These response functions were obtained by fitting several models under (2) that will later be described in detail; response style factors were allowed to correlate with substantive traits and each other, scoring functions were fixed, item slopes were estimated, and item parameter estimates appear in Table 1. In the absence of response styles, if this item loads on a hedonic benefits factor whose scoring functions are ordered as in the GPC model, it may result in response functions similar to those that appear in the top row of Figure 2. If ERS also affects this item and is thought to be a preference for the endpoint categories, defined by scoring functions of $[1 \quad 0 \quad 0 \quad 0 \quad 1]$, this results in response functions in the first column in Figure 2. Thus, Figure 2 depicts cross sections of response functions with the x-axis representing hedonic benefits, and each row representing a different level of the response style. The top row in Figure 2 is at the mean of the response style (ERS=0), the middle row is above the mean (ERS=1), and the bottom row is below (ERS=-1). Notice that when ERS=1, the response categories appear contracted towards the middle of the scale and the response functions for the lowest and highest categories have high probabilities for a large range of the hedonic benefits latent trait. Conceptually, if a respondent is high on ERS, they are more likely to pick one of the endpoint categories and we are less confident that such a response indicates that they are actually high (or low) on hedonic benefits. In contrast, when ERS=-1, the middle three categories are exaggerated and have higher probabilities across a much wider range of hedonic benefits.

The remaining two columns in Figure 2 are also from models with a single response style factor and represent response function cross-sections where MRS is defined as pref-

erence for the middle category with scoring functions $[0 \quad 0 \quad 1 \quad 0 \quad 0]$ and EMRS as a graded preference for the middle versus endpoint categories defined by scoring functions $[2 \quad 1 \quad 0 \quad 1 \quad 2]$. These examples illustrate the possibility that a fine distinction between ERS, MRS, and EMRS is possible, with the modeling of each resulting in different response function shapes. For example, high on MRS (MRS=1) results in a middle category that has a relatively high probability across hedonic benefits, which is not completely redundant with being low on ERS (ERS=-1). In addition, note the difference between ERS=1 and EMRS=1, suggesting that an exclusive preference for the endpoint categories (ERS) is not the same as graded preference for the endpoints versus categories that are towards the middle of the scale (EMRS).

The example response functions in Figure 3 include two correlated response style factors, ERS and MRS, with scoring functions defined as before. These examples illustrate additional response function shapes not possible when only a single response style dimension is in the model. For example, being high on both of these factors (ERS=1, MRS=1) is indicative of more frequent use of the endpoint categories *and* the midpoint category whereas being low on these factors (ERS=-1, MRS=-1) indicates more frequent use of the intermediate categories.

Although we have focused on ERS, MRS, and EMRS, the scoring functions simply define how the order of the categories relate to each latent trait and may follow almost any user-specified values. Table 2 lists several additional options for defining scoring functions. A useful heuristic is to consider scoring functions analogous to contrasts used for categorical predictors in linear models (e.g., regression, ANOVA). For instance, we could have used the contrast $[1 \quad -1 \quad -1 \quad -1 \quad 1]$ for ERS, which would have yielded an equivalent item model, but changes the scaling of the item slope. A positive slope for ERS indicates that selection of the endpoint categories would reflect being higher on the ERS latent trait and that both endpoints are weighted equally. An additional heuristic is to consider the scoring functions analogous to category weights used in ad-hoc sum

score approaches to measuring response styles (for a review, see de Jong, Steenkamp, Fox, & Baumgartner, 2008). For example, counting the number of times a participant selects the endpoint categories as an index for ERS conceptually maps onto the scoring function values used in Table 2, yet does not readily incorporate a measurement model to disentangle style from content.

Our approach also retains the possibility of modeling ARS by defining the scoring function for an orthogonal ARS dimension as in the GPC model and fixing ARS loadings equal across items (Maydeu-Olivares & Coffman, 2006; Savalei & Falk, 2014), *or* defining ARS as the choice of categories above the midpoint of the response options, $[0 \quad 0 \quad 0 \quad 1 \quad 1]$. However, we agree with Maydeu-Olivares and Coffman (2006) that whether the emerging dimension represents ARS or something substantive must be determined based on theoretical grounds.[1] In general, the interpretation of response style factors may be more difficult as the scoring functions more closely resemble possible substantive traits. For instance, in some scoring algorithms for SDR (Paulhus, 1991), the top two-most categories on 7-point positively keyed items count towards SDR, possibly representing the scoring function $[0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]$. While it is possible to fit a model with this SDR definition and a substantive dimension simultaneously, additional work is needed to establish the validity and interpretation of the emerging dimensions.

We also note that *the scoring functions need not be fixed to the exact same values across items for any given dimension*. If there is sufficient theory or evidence that a particular category for a particular item defines a response style, then researchers may define the scoring function for that item in a unique way that contrasts that category with the other available response options (e.g., see Table 2). For example, if the 4th category is the most socially desirable response for item 1, but the 3rd category is the most desirable for item 2 (e.g., Kuncel & Tellegen, 2009), these two items may have *different* scoring functions for

[1] We refer the reader for a more extensive discussion of the former definition of ARS in Maydeu-Olivares and Coffman (2006). We also suspect that in some cases its interpretation as ARS may be bolstered by low correlations with substantive dimensions, the presence of both positive and negatively worded items, or substantive traits that are nearly uncorrelated.

the *same* SDR latent trait. Yet, the scoring functions also need not follow integer values; for example, participant or expert ratings of the social desirability of different response options may determine the category weights (see last row of Table 2).

3.3 Additional Details

The models in (1) and (2) are not identified without additional constraints as discussed in detail for the unidimensional case by multiple authors (Bock, 1972; Thissen et al., 2010; Thissen & Cai, in press; Thissen & Steinberg, 1986) and for the multidimensional case by Johnson and Bolt (2010). One indeterminacy common to multidimensional factor analytic models concerns rotation and reflection of factor slopes, and is discussed in further detail by Johnson and Bolt (2010). In short, exploratory item factor analysis will require an additional $D(D-1)/2$ constraints on category or item slopes.

In the present application, we are utilizing the model in a confirmatory mode, which entails additional constraints. If we include only substantive latent traits and items load on only one factor, such a simple structure allows for identification and estimation of the correlation between factors. The addition of response style factors introduces additional complexity as items load on more than one factor and we may wish to estimate the correlation between response style and substantive factors. Models such as those depicted in Figure 1 may not intuitively appear to be identified, but may be provided that the scoring functions for response style factors are not redundant with each other and that of the substantive factors. In Supplementary Materials we provide an example of how to check the identifiability of a model and test it on a hypothetical example with one substantive trait and two response style traits (ERS and MRS) that are all correlated and load on all items.[2] That said, estimation problems may be fewer if slopes on the response style factor are equal or fixed across items (e.g., Bolt & Newton, 2011; Johnson & Bolt, 2010), or if response style factors are kept orthogonal to substantive dimensions (e.g., theoretically consistent with Couch & Keniston, 1960; Jackson & Messick, 1961).

[2]This process involves checking the rank of the Jacobian matrix formed by differentiating each model implied probability for all possible response patterns with respect to each model parameter.

Another indeterminacy is inherent to the translation invariance of the multinomial logistic response functions. The addition of a constant to $\mathbf{c}$ or any given set of slope parameters for a particular dimension, $\tilde{\mathbf{a}}_k$, will yield the same value for the response function. Constraints often include either the fixing of one intercept parameter and one slope per dimension (e.g., $c_1 = 0$ and $\tilde{a}_{d,1} = 0$) or constraints that enforce $\sum_{k=1}^{K} \tilde{a}_{d,k} = 0$ for all d and $\sum_{k=1}^{K} c_k = 0$.

Under the reparameterization in (2) as under some versions of the original nominal model, constraints necessary for identification are formed via products of the actual item parameters with other constant contrast matrices, $\mathbf{T}$ (Thissen et al., 2010; Thissen & Cai, in press). For instance, $\mathbf{c}$ involves $K-1$ parameters in γ, and row d of $\mathbf{S}$ involves $K-1$ parameters in $\boldsymbol{\alpha}_d$. Details of these parameters and contrast matrices are discussed in Appendix A.

3.4 Relationship to Similar Approaches

To our knowledge, the reparameterization in (2) has not been previously used to model response styles, but can subsume models used by Bolt and colleagues (Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010). To aid in comparing various submodels, Table 3 presents how slopes and scoring functions for a single dimension and intercepts may be defined (subject to identification constraints on scoring functions and intercepts; see also Appendix A). The two features that distinguish our approach are the added ability to estimate overall item slopes for response style dimensions that differ across items, and user-defined response style scoring functions that may differ across items (see "Response style" in Table 3). The item models used for response styles by Bolt and colleagues can accommodate category slopes representing the order of categories for a given latent trait (e.g., $\tilde{\mathbf{a}}$) that are fully estimated from the data (subject to identification constraints) and different across items (NRM-free), estimated from the data but constrained equal across items (NRM-equal), or all fixed to prespecified values (NRM-fixed; often with the variance of the latent trait estimated). Under NRM-free

and NRM-equal, there is no guarantee that the estimated order of categories for a given factor will neatly represent any known response style. If the order of categories for a factor is fixed or constrained equal across items as under NRM-fixed and NRM-equal, an assumption is being made that the factor affects all items equally. In the parameterization in (2) scoring functions are separated from overall item slopes, which allows for fitting all of the aforementioned modeling approaches as well as allowing item slopes for response style dimensions to vary across items. We now briefly describe how the models fit by Bolt and colleagues can also be accommodated under (2).

Based on the results of an exploratory item factor analysis, Bolt and Johnson (2009) fit a two-dimensional model with orthogonal factors in which the category slopes were estimated from the data, but were equal across items for the response style factor and free for the substantive factor (NRM-equal and NRM-free) or constrained for both factors (both NRM-equal). NRM-free for a factor is simply the full nominal model for that dimension (e.g., full rank $\mathbf{T}$) and requires estimating a and $\alpha_2, \ldots, \alpha_{K-1}$ parameters for that dimension. To fit NRM-equal under (2), these same parameters are estimated, but are constrained equal across items (see Table 3).

To data that theoretically contained two correlated substantive dimensions, Bolt and Newton (2011) fit three models. After fitting a 2-dimensional PC model for substantive factors only, a second model included an additional NRM-free factor. A third, and best fitting model additionally used NRM-fixed with slopes of $[1 \quad -1 \quad -1 \quad 1]$, and estimated the factor variance-covariance matrix. Fitting the PC or NRM-fixed approaches both involve fixing the scoring functions to predetermined values by fixing $\alpha_1 = 1$ and $\alpha_2, \ldots, \alpha_{K-1} = 0$ and setting the first column of $\mathbf{T}$ to the desired scoring function (see also Appendix A). In addition, item slopes are constrained equal across items *or* fixed to a predetermined value (e.g., 1) while estimating of the variance of the latent trait.

Finally, Johnson and Bolt (2010) presented a two-factor model used in simulations in which the substantive dimension (and intercepts) followed a "stereotype" (ST) model

and a second dimension was NRM-equal. A second more constrained model was fit to actual data in which the substantive dimension (and intercepts) followed the rating scale (RTS) model and the second dimension was again NRM-equal. In both models the correlation between the response style factor and substantive factor was estimated. Although only used for substantive dimensions under Johnson and Bolt (2010), we note that item slopes for the ST model are decomposed into overall item and category effects. Under our parameterization, this is equivalent to freely estimating a, but constraining $\alpha_2, \ldots, \alpha_{K-1}$ or the scoring functions for a given dimension to be equal across items (see also Table 3). The RTS model has item slopes parameterized in the same way as the PC model. However, both the RTS and ST models are characterized by intercepts that are separated into overall item and category effects (i.e., $c_{jk} = \beta_j + c_k$ with c_k equal across items) and two constraints for identification.[3] Such intercept constraints are possible by using those described by Thissen and Steinberg (1986) for the RTS model.

On the basis of the above review of previous works (Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010), we point out that category slopes were often either fully exploratory across items (NRM-free) or constrained equal across items (NRM-equal, NRM-fixed, PC, RTS). This observation holds for both substantive and response style dimensions with the lone exception the ST model for a single substantive dimension for a model in simulations by Johnson and Bolt (2010). Although response styles are often thought to affect all items equally, such an assumption may be questioned on empirical and theoretical grounds (e.g., Ferrando, Lorenzo-Seva, & Chico, 2003; Krosnick, 1991; Trott & Jackson, 1967), and these possibilities omit the case where a researcher may wish to explicitly test this assumption. In the case of substantive dimensions, this may result in an unrealistic assumption that the items are equally related to the latent trait. The alternative allowed by the reparameterization of the MNRM is for researchers to determine the order of each item's categories and how they relate to each latent trait by

[3]Johnson and Bolt (2010) constrain one c and one β parameter to 0.

fixing the scoring functions, while allowing estimation of overall item slopes. This may ease the interpretation of how response styles and substantive dimensions are thought to affect item responses and provide a compromise between exploratory and highly constrained parameterizations. Although the ST model decomposes slopes into overall item and category slopes which may allow for this possibility, it does not allow for category slopes that define a particular response style in different ways across items, and also includes additional constraints on intercepts that may not be desired.

4 Estimation, Standard Errors, and Scoring

Full-information maximum marginal likelihood estimation is widely used in item parameter estimation (Bock & Lieberman, 1970; Bock & Aitkin, 1981; Schilling & Bock, 2005). To avoid a prohibitive number of quadrature points necessary for integration, we use the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm, which is described in further detail elsewhere (Cai, 2010a, 2010b; Monroe & Cai, 2014; Yang & Cai, 2014). MH-RM is a stochastic approximation method that is able to handle a high number of latent traits and converges to the maximum marginal likelihood solution. We briefly describe MH-RM, but refer the reader to additional details in Appendix B and C and the above references.

Recall $\mathbf{X}$ is an $N \times D$ matrix of factor scores and $\mathbf{Y}$ is an $N \times n$ matrix of item responses. When one takes the view that the latent factor scores $\mathbf{X}$ are the missing data and the item responses $\mathbf{Y}$ are the observed data, MH-RM proceeds by "filling in" the missing data and iteratively updates the parameters of the complete-data model until convergence. In MH-RM, the imputations for $\mathbf{X}$ are drawn from the posterior predictive distribution, $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$, which is proportional to the complete-data log-likelihood and where $\boldsymbol{\omega}$ contains all model parameters.

At iteration $(p+1)$ of the algorithm and given initial parameter estimates, m_p imputations of the latent traits $\left\{\mathbf{X}_t^{(p+1)}; t = 1, \ldots, m_p\right\}$ are combined with the observed data, $\mathbf{Y}$, to represent m_p complete data sets. Typically $m_p = 1$ imputation per iteration is

sufficient. For this, we use the same Metropolis-within-Gibbs sampler as in Cai (2010a).

Once imputations from $\Pi(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})$ for iteration $p+1$ are obtained, they are used to provide an approximation for the observed data gradient of the marginal log-likelihood, $\dot{l}(\boldsymbol{\omega}|\mathbf{Y}) = \frac{\partial l(\boldsymbol{\omega}|\mathbf{Y})}{\partial \boldsymbol{\omega}}$. In particular, the complete data first-order derivatives, $\dot{l}(\boldsymbol{\omega}|\mathbf{X}_t^{(p+1)},\mathbf{Y}) = \frac{\partial l(\boldsymbol{\omega}|\mathbf{X}_t^{(p+1)},\mathbf{Y})}{\partial \boldsymbol{\omega}}$, for each imputation are computed and averaged to form $\mathbf{g}_{p+1}$. This approximation of $\dot{l}(\omega|\mathbf{Y})$ works because due to Fisher's identity (Fisher, 1925), $\dot{l}(\boldsymbol{\omega}|\mathbf{Y})$ is the same as the expectation of the gradient vector for the complete data log-likelihood, $\dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})$, over the posterior predictive distribution of the latent traits, $\Pi(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})$. Directional information for updated parameter estimates is provided by $\mathbf{g}$, and additional curvature information is computed via a recursive approximation to the conditional expectation of the complete data information matrix, $\boldsymbol{\Gamma}$ (see Appendix B and C). Parameter estimates are then updated by:

$$\boldsymbol{\omega}^{(p+1)} = \boldsymbol{\omega}^{(p)} + \beta_p \boldsymbol{\Gamma}_{p+1}^{-1} \mathbf{g}_{p+1} \tag{3}$$

where β_p is a gain constant. After updating parameter estimates, the next iteration begins with updated imputations from the posterior predictive distribution.

The sequence of gain constants, $\{\beta_p; p \geq 1\}$, can be defined to slowly decay across iterations. This constitutes the Robbins-Monro part of the algorithm and filters out noise across iterations. In practice, recent instantiations of MH-RM (Cai, 2013; Monroe & Cai, 2014; Yang & Cai, 2014) divide iterations into several stages that have different gain constants and/or starting values. Stage 1 is often intended to improve parameter estimates that may be far away from the MLE and has a gain constant that remains the same across iterations. Once parameter estimates stabilize around a neighborhood of the MLE, Stage 2 iterations begin. They also have the same gain constant across iterations and are often intended to obtain good starting values for the next stage. In Stage 3, the mean of parameter estimates across iterations for Stage 2 are often taken as starting

values for Stage 3, which has gain constants defined by $\beta_p = \beta_0/p^{\varepsilon}$ with $1/2 < \varepsilon \leq 1$ and β_0. Iterations may be terminated after the minimum change in parameter estimates is below some threshold (e.g., 1e-4) for across a window of previous iterations (e.g., 3).

The MH-RM algorithm proceeds in nearly the exact same way in multiple group analyses: Imputations from the posterior predictive distribution can be conducted separately for each group, as are computations of complete data gradient and information matrices. The complete data log-likelihood is simply the sum of each group's log-likelihood. Similarly, the observed data gradient, $\dot{l}(\boldsymbol{\omega}|\mathbf{Y})$, at iteration $p+1$ can be approximated by combining $\mathbf{g}$ from each of the groups (see Appendix C for further details).

For the present application, equality constraints on parameters may be implemented in the same way as in (Cai, 2010b). Specifically, linear constraints can be implemented by $\boldsymbol{\omega}_r = \mathbf{c} + \mathbf{L}\boldsymbol{\omega}$, where $\boldsymbol{\omega}_r$ is the $q \times 1$ vector of parameters that satisfies the constraints, $\boldsymbol{\omega}$ represents all free parameters ($p \times 1$ with $p \leq q$), $\mathbf{c}$ is a $q \times 1$ vector of constants, and $\mathbf{L}$ is a $q \times p$ matrix that implements linear (equality) constraints. The gradient and information matrix for updating $\boldsymbol{\omega}$ is then obtained by $\mathbf{g} = \mathbf{g}_r\mathbf{L}$ and $\boldsymbol{\Gamma} = \mathbf{L}'\boldsymbol{\Gamma}_r\mathbf{L}$, where $\mathbf{g}_r$ and $\boldsymbol{\Gamma}_r$ involve derivatives with respect to $\boldsymbol{\omega}_r$, ignoring the presence of constraints.

MH-RM requires reasonable starting values at iteration 0 for draws of the latent traits, $\mathbf{X}^{(0)}$, from the posterior predictive distribution. Under the MNRM we computed the following weighted mean scores for each participant based on starting values for item slopes and scoring functions, $\mathbf{x}_i^{(0)} = \frac{1}{n_i}\sum_{j=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(y_{ij})[\mathbf{a}_j \circ \mathbf{s}_{k,j}]'$ where j item indexes $\mathbf{a}$ and $\mathbf{s}_k$, and $\mathbf{X}^{(0)}$ was then normalized across participants. In the analyses reported in this paper, unit weights were used for all free $\mathbf{a}$ starting values.

Standard errors under MH-RM are often obtained following a formula given by Louis (1982), whose elements can be recursively approximated during optimization (Cai, 2010a). As noted by others (Yang & Cai, 2014), sometimes this approach may not yield valid standard errors (i.e., negative diagonal elements of the inverse information matrix), possibly due to stability issues for the second-order derivative approximations when the

fraction of missing information for the latent variable model is high, e.g., when the dimensionality is large or model structure complex. In such a case, Monte Carlo integration upon convergence of the MH-RM algorithm can be used in a more direct application of Louis's (1982) formula: Parameter estimates are fixed at the MLE and draws of the latent traits are taken from the posterior predictive distribution (see Diebolt & Ip, 1996; Fox, 2003). Following Yang and Cai (2014) we refer to the former approach as *recursively approximated standard errors* and the latter as *post-convergence standard errors*. Both are studied in our simulations.

Finally, upon estimation of item parameters, we can obtain factor scores of individual respondents using the *expected a posteriori* (EAP) method (Bock & Mislevy, 1982). Although typical applications of EAP use quadrature to integrate across the distribution of latent traits, Monte Carlo integration is also possible by taking a large number of draws from the distribution of latent traits. We utilize the latter approach in empirical examples and simulations.

5 Empirical Demonstration

Returning to our PROMIS®example, we fit a series of models listed in Table 4 to daily and non-daily smokers simultaneously.[4] The mean and variances of the latent variables were fixed for daily smokers (all $\mu = 0$ and all $\sigma^2 = 1$; the reference group), but estimated for non-daily smokers (the focal group). Additionally, all factor covariances were freely estimated for both groups, including those involving response style factors. Item parameters (intercepts, and substantive and response style slopes) for 78 common items were constrained equal across groups. For simplicity, these anchors were chosen based on differential item functioning analysis undertaken as a part of the initial calibration results for the PROMIS®smoking initiative, though we note that multidimensional DIF testing incorporating response style factors could also be undertaken (e.g., Bolt & John-

[4]All programming for empirical examples and simulations was done by the first author and implemented in R (R Core Team, 2012) with first- and second-order analytical derivatives using the Rcpp (Eddelbuettel & Francois, 2011) and RcppArmadillo packages (Eddelbuettel & Sanderson, 2014). Models 1, 4, 6, and 8 are used for the example response functions in Figures 2 and 3.

son, 2009). If one is not interested in comparing response style means across groups, all slopes for the response style factors could also be unconstrained across groups. Confidence intervals for AIC and BIC were based on Monte Carlo integration using 25,000 points upon model convergence (see Appendix C). Estimation details typically included a burn-in of 10 for draws from the posterior predictive distribution, followed by 2,000 Stage 1 iterations, 500 Stage 2 iterations with $\beta_0 = .75$, $\varepsilon = 1$, a convergence window of 3 with a $5e-5$ tolerance, and 1 imputation per iteration of the algorithm. All models converged within 8,500 Stage 3 iterations. In some cases, a few steps (e.g., 5-100) with tiny gain constants (e.g., 5e-5 to .01) were necessary to obtain better starting values before commencing with Stage 1 iterations, or use of starting values based on estimates from a more constrained model.[5]

Models 1-11 pertain to those we allude to in Section 2, whereas Models 12-17 represent models similar to those used by Bolt and colleagues in Section 3.4. We use the abbreviations 6D and 6D ST to refer to modeling of the six substantive dimensions using the GPC or ST models, respectively. The remaining dimensions include free estimation of the order of categories for a particular dimension (NRM) or response style factors ERS, MRS, and EMRS using scoring functions appearing in Table 2. The column for **a** in Table 2 indicates whether item slopes were free or constrained equal across items for the substantive and each of the response style dimensions. The first 11 Models therefore include ERS, MRS, and EMRS separately, and the combination of ERS and MRS with all possible permutations of fixing response style slopes equal or setting free across items.

AIC and BIC both indicated that the best fitting model was one that included both ERS and MRS, modeled as separate factors and with freely estimating ERS slopes. BIC clearly suggested that MRS slopes should be further fixed equal across items (Model 10; see Table 4), whereas AIC did not clearly differentiate between this model and uncon-

[5]A previous version of the paper included Models 2-11 in which response style factors were orthogonal to each other and all other dimensions. Use of starting values from these models aided in estimating Models 2-11 in the current paper, in which the covariances among all factors are estimated. Results from these and other fitted models appear in Supplementary Materials.

strained slopes (Model 8). We are inclined to select Model 10 as BIC clearly prefers this model (AIC yielded an inconclusive decision) and tends to select more parsimonious models than does AIC. That these models were preferred to those including EMRS suggests that ERS and MRS may not be different poles of the same dimension, and that the assumption that response style factors affect all items equally may not be tenable for ERS, but may be for MRS. In fact, estimated factor covariances for Model 10 (Table 5) reveal that ERS and MRS were nearly uncorrelated with each other and most substantive dimensions. It is notable that either Model 8 or 10 exhibited better fit than all of Models 12-17.

Models 12-15 are similar to those fit by Bolt and Newton (2011) in which both substantive and response style slopes were fixed across items for any given dimension (NRM-fixed). These parallel Models 3, 5, 7, and 11 in which there are free substantive slopes. AIC and BIC both preferred Models 3, 5, 7, and 11, suggesting that constraints on substantive slopes may be unrealistic. Model 16 is similar to that fit by Bolt and Johnson (2009), in which an additional NRM-equal factor is added to the six-dimensional model. This Model's fit is comparable to Models 2 and 3 (that also model ERS), and the estimated category slopes for this exploratory dimension are suggestive of ERS or EMRS, $[0 \quad -.93 \quad -1.10 \quad -.89 \quad .27]$, though does not provide better fit than any models that have both ERS and MRS as separate dimensions. Model 17 is similar to a model fit by Johnson and Bolt (2010) and parallels Model 16 with the addition of intercept constraints imposed by the ST model, which appear to hinder model fit.[6]

To inspect changes in subject scores on the substantive traits, we conducted EAP scoring using Monte Carlo integration with 100,000 draws per subject under Models 1 (substantive traits only) and 10 (ERS and MRS with MRS slopes constrained equal). EAP scores for daily smokers for each substantive trait under these Models are plotted in

[6]The ST model without response styles and additional variants of Models 16 and 17 (without substantive-response style factor correlations) appear in Supplementary Materials and did not provide any improvements in model fit.

Figure 4, with ERS and scores from Model 10 used to group subjects into low (< -1), moderate (-1 to 1) and high (> 1) on the response style. Overall there is a strong, but imperfect relationship between scores under each model, suggesting that inclusion of response styles in Model 10 (relative to their absence from Model 1) results in adjustments to subject's scores. Note that in the top row of Figure 4, subjects high in ERS tended to have their scores adjusted towards the mean (indicated by the slope of the cloud of points being slightly flat), whereas in the lowest panel those with low ERS scores tended to have their scores adjusted away from the mean on the substantive traits (indicated by a steeper slope in the cloud of points). A consistent pattern of score changes due to MRS is less obvious and appears in Supplementary Materials.

6 Simulations

Two small sets of simulations were conducted to check the ability of the model and estimation approach at recovering model parameters, subjects' factor scores, and estimating standard errors. A secondary goal of Simulation 1 was to check the consequences of a misspecified model (e.g., failing to estimate response style factors when included in the data generation model, and vice versa). Since it is commonly thought that it is easier to disentangle response styles from substantive traits that are unrelated (e.g., Greenleaf, 1992), a secondary goal of Simulation 2 was to test for differences in the outcomes if the data generating model included substantive factors that were either nearly uncorrelated or moderately correlated.

The data generating models for all conditions were based in part on models fit to actual data: 16 hedonic benefits (HB) and 27 nicotine dependence (ND) items from PROMIS®daily smokers. In all cases, the model included two correlated substantive factors for these two dimensions and differed in terms of included response style dimensions. In all but one case later described, item parameter estimates and factor correlations were then treated as population values for data generation (see Supplementary Materials for individual values). For each data generating condition, one-hundred datasets of

$N = 2000$ were created assuming multivariate normal latent traits.

Details common to estimated models in simulations included a burn-in of 10 for draws from the posterior predictive distribution, followed by 300 Stage 1 iterations, 500 Stage 2 iterations with constant gain constants of .5, an initial Stage 3 gain constant of $\beta_0 = .1$, $\varepsilon = .75$, a convergence window of 3 with a $5e-5$ tolerance, and 1 imputation per iteration with thinning of 5 between draws. Instead of true parameter values, unit slopes, zero intercepts, and zero factor correlations were used for all starting values. All models converged in under 5,100 Stage 3 iterations. Post-convergence standard errors were computed when the true model was fit to the data with a minimum of 100 draws and yielded valid standard errors in 100% of cases in 310 iterations or less. For each fitted model, EAP scores for all factors were computed using 25,000 Monte Carlo draws.

6.1 Simulation 1

6.1.1 Method

Three data generating models were used: 1) a two-factor substantive traits only model (GPC), 2) a four-factor model with ERS and MRS factors that were orthogonal to each other and the substantive traits (RS), and 3) a four-factor model in which ERS and MRS were correlated with each other and the substantive dimensions (RS-cor). Scoring functions for ERS and MRS were $[1 \quad 0 \quad 0 \quad 0 \quad 1]$ and $[0 \quad 0 \quad 1 \quad 0 \quad 0]$, respectively. To each generated dataset, we fit three models: GPC, RS, and RS-cor. Thus, for each data generating condition, the true model was always fit to the data as well as two overfitted or misspecified models. AIC and BIC point estimates were also calculated for each estimated model based on 25,000 Monte Carlo integration points.

6.1.2 Results

When the true model was fit to the data, no appreciable estimation bias was observed for any type of parameter (see top row of Figure 5). Specifically, raw bias across replications averaged to .01 or less in absolute value for each type of model parameter. Root mean square error (RMSE) was similarly good. For instance, under all three

data generating models average RMSE (across items) was .06 for substantive slopes (all $SD = .02$), and was .06 and .07 for γ parameters (all $SD = .04$) for the GPC and both response style models, respectively. RMSE for response style slopes was nearly as good as those for substantive traits, averaging .08 to .09 for ERS for the RS ($SD = .01$) and RS-cor ($SD = .02$) models, respectively, and was .08 for MRS for both models (both $SD = .01$). Finally, RMSE for factor covariances was slightly smaller for the GPC (.02) and RS models (.02) than for the RS-cor model ($M = .03$, $SD = .005$).

Recursive standard errors were available in 94%, 87%, and 84% of cases for the GPC, RS, and RS-cor models, respectively, with the remaining cases having at least one negative value on the diagonal of the inverse information matrix. In what follows, we report on recursive standard errors from valid replications and post-convergence standard errors from all replications. The middle and bottom row of Figure 5 compare empirical standard deviations against recursive and post-convergence standard errors, respectively. Both types of standard errors provided a reasonable approximation to empirical standard deviations. The mean of standard error estimates across replications for each data generating conditions were mostly (68% or more) within .01 of the empirical standard deviations and all were within .05. The quality of standard errors declined slightly with the more complex model (RS-cor) and with post-convergence standard errors. For example, for all other conditions the maximum discrepancy was less than .04, and 77% or more of mean standard errors were within .01 of empirical standard deviations.

Finally, we examined trait recovery (Table 6) using both Pearson correlations and $\text{RMSE}_\theta = \left(N^{-1}\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2\right)^{1/2} \times 100$, where N is the number of respondents, $\hat{\theta}_i$ is the EAP score and θ is the true trait score. Fitting the true model to data resulted in good substantive trait recovery, with r ranging from .96 to .98 and RMSE_θ from 18.24 to 29.65. Recovery of ERS was also decent, $r = .91, .91$, and $\text{RMSE}_\theta = 41.49, 40.74$ for RS and RS-cor true models, respectively, but less impressive for MRS, $r = .77, .78$, and $\text{RMSE}_\theta = 63.80, 62.85$. Overfitted models did not have worse trait recovery. For

instance, trait recovery for substantive factors under the GPC data generating model was excellent when fitting the true model, $r = .97, .98$, and $\text{RMSE}_\theta = 24.48, 18.24$, but was nearly indistinguishable from when the RS or RS-cor models were fit to the same data. Similarly, fitting the RS-cor model to RS data also yielded good trait recovery. Trait recovery worsened when the GPC model was fit to data generated with response styles. For example, for RS-cor data, substantive trait recovery dropped to $r = .91, .93$, and $\text{RMSE}_\theta = 41.14, 36.41$ for the GPC fitted model. However, the consequences of fitting RS to RS-cor data were negligible, suggesting that perhaps response style factor correlations (.31 or less) from the data generating model were not large enough to have a substantial effect on RS model results (see also Supplementary Materials).

Although the study was not designed to test the efficacy of AIC and BIC in model selection, it is worth noting that BIC selected the true model in 100% of cases. AIC selected the correct model in 99% of cases when GPC was the true model, and 100% of cases when RS-cor was the true model. When RS was the true model, AIC always preferred the response style models over the GPC model, but only selected the RS model over the RS-cor model in 52% of cases, thus sometimes suggesting unnecessarily that response style factors be correlated with other factors. However, as seen with trait recovery, the consequences of this overfitting may not be harmful.

6.2 Simulation 2

6.2.1 Method

Two data generating models were used and differed on whether the substantive traits were moderately correlated (Mod-cor) or weakly correlated (Low-cor). In both cases, a correlated four-factor model with ERS and ARS was used. ARS was conceptualized as choosing a score above the midpoint of the Likert scale, with scoring function $[0 \quad 0 \quad 0 \quad 1 \quad 1]$. This allowed it to correlate with substantive dimensions, yet based on the PROMIS®data the estimated factor correlations of ARS with substantive dimensions were less than .1 in absolute value (see Supplementary Materials). Since the estimated

HB-ND correlation was large, $r = .46$, item parameters and the factor correlation from this model served as the Mod-cor data generating model. For the Low-cor model, the same item parameters and factor correlations were used, except the HB-ND correlation was arbitrarily set to .05. To each generated dataset, only the true model was fit.

6.2.2 Results

On average, zero bias was observed for both models and all parameters (see top row of Figure 6). RMSE for Mod-cor ranged from .02 to .13 ($M = .08$, $SD = .02$) for item slopes, .02 to .19 ($M = .07$, $SD = .04$) for γ parameters, and from .02 to .04 ($M = .03$, $SD = .01$) for factor correlations. These values differed little from those for the Low-cor model, from .02 to .18 ($M = .08$, $SD = .02$) for item slopes, .02 to .22 ($M = .07$, $SD = .04$) for item intercepts, and from .02 to .03 ($M = .03$, $SD = .00$) for factor correlations. Similar to the previous set of simulations, under both models RMSE for substantive slopes (both $M = .06$, $SD = .02$) was slightly better than that for response style slopes. Specifically, RMSE for ERS was about the same across the Mod-cor ($M = .09$, $SD = .01$) and Low-cor ($M = .09$, $SD = .02$) models, and similar to that for ARS under the Mod-cor ($M = .09$, $SD = .01$) and Low-cor ($M = .10$, $SD = .02$) models.

For both models, recursive standard errors were available for 84% of replications. Both recursive and post-convergence standard errors were well calibrated, with a possible exception for post-convergence standard errors for the Mod-cor model (see row 3 of Figure 6), which on average were underestimated by .01 relative to empirical standard deviations. Even in this condition, however, the maximum difference between average estimated standard errors and empirical standard deviations was .08, and 57% of mean standard errors were within .01 of empirical standard deviations. In all other cases, the maximum difference was always under .04 and 79% or more were within .01.

Recovery of substantive traits was good for the Mod-cor model, $r = .94, .97$ and $\text{RMSE}_\theta = 33.18, 26.17$, (for HB and ND, respectively) and the Low-cor model, $r = .94, .97$ and $\text{RMSE}_\theta = 33.12, 25.86$. Recovery of ERS was also good for Mod-cor $r = .91,$

$\text{RMSE}_\theta = 41.20$, and Low-cor, $r = .91$, $\text{RMSE}_\theta = 40.79$. Recovery of ARS was slightly less impressive for both Mod-cor, $r = .80$, $\text{RMSE}_\theta = 60.27$, and Low-cor, $r = .80$, $\text{RMSE}_\theta = 59.67$. Pearson correlations for trait recovery were identical to the second decimal place for each type of model, and RMSE values were trivially smaller for the Low-cor model.

7 Discussion

We have presented a flexible multidimensional item response theory model using a novel MNRM parameterization that is able to model multiple response styles across multiple substantive constructs. The model is able to subsume previous approaches (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Johnson & Bolt, 2010) while allowing for item slopes for response style factors to vary across items, and definition of response styles or other factors whose order of categories is *different* across items.

Results from the empirical application suggest that our preferred modeling approach has the potential to fit data better than previous uses of the MNRM for modeling response styles. For instance, allowing response style and substantive item slopes to be free across items often resulted in better model fit, and is possible even when response style factors are correlated with other dimensions. That our model allows for estimating varying slopes should open the door to extensions or additional research for identifying item features most likely to yield low/high response style slopes. While we must be cautious not to overgeneralize from a single illustrative example in which the study design was not initially intended to investigate response styles, the example also illuminated the possibility that ERS and MRS may be separate factors in that the best fitting models had these factors as separate dimensions that were nearly uncorrelated. This builds upon previous research using the MNRM where a single response style dimension reveals an ERS or EMRS factor (Bolt & Johnson, 2009; Johnson & Bolt, 2010; Kieruj & Moors, 2010, 2013), or a factor in which preference is for the endpoints or the midpoint (Moors, 2004; Morren et al., 2011). We are at the mercy of AIC and BIC, however, and do not know with complete certainty which model is a better or more useful approximation to reality

and whether this result will generalize. For instance, with a longer Likert scale (e.g., 7-point), it is possible that extreme response style may better resemble a graded preference for the endpoints (like EMRS) that could be simultaneously modeled with MRS.[7]

The simulations also showed that our estimation and scoring approach is feasible and results in good parameter recovery and well-calibrated standard errors. Simulation 1 additionally suggested that it may be prudent to err on the side of modeling response styles as the consequences of overfitting are outweighed by the consequences of failing to include response styles. For Simulation 2, while it may be possible that low correlations among substantive factors are better for disentangling response styles from substantive traits (and ARS in particular), we found little evidence of such an effect under the limited set of data generation conditions studied. However, we consider these results preliminary and better examined by a fuller set of simulations and real data analyses.

Our modeling framework is general enough to accommodate arbitrarily defined response styles, provided that the response styles are thought to affect item responses in a compensatory manner. While we focused on ERS and MRS in this paper's empirical example, we also specifically outlined how our approach may be adapted for more refined distinctions among response styles (e.g., EMRS) and other arbitrarily defined response styles including ARS and SDR. Future research may be undertaken to demonstrate modeling of such additional response styles. Yet we note that our model's flexibility should not be a substitute for good scale construction practices and strong theory. Under conditions of low information (e.g., few good items, subjects) it may still be difficult to disentangle response styles from substantive constructs and additional research may identify cases where this is true. This issue or difficulty interpreting the response styles factor may occur, for example, with response style constructs whose scoring functions are similar to substantive traits (e.g., ARS) or when response style factors, such as SDR, are highly correlated with some substantive trait. There will also be an upper limit to

[7]We thank an anonymous reviewer for suggesting how extreme response style may be manifested in items with more response options.

the number of factors that can be simultaneously estimated from the data, and different configurations will result in models with different substantive meanings.

Likewise, our approach has also not yet been thoroughly compared with alternative modeling approaches, yet we note that there are few existing comparisons among other approaches. For example, it is an open question as to whether the compensatory effect of response styles follows the process modeled with the MNRM, or if completing Likert-type items is closer to a multiple-stage decision making process (Böckenholt, 2012; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013). Our approach also uses the same item data for estimation of response styles, whereas it is also possible to use data external to the constructs of interest (Bolt, Lu, & Kim, in press) or measurement instruments specifically designed for this purpose (Paulhus, 1991). Therefore, while our current modeling framework showed promise in an empirical application and simulation study, there is a growing need for additional comparisons among full-information approaches with both real and simulated data.

Appendix A. Contrast Matrices

Contrast matrices facilitate the implementation of the identification constraints. In particular, $\mathbf{c} = \mathbf{T}_c\gamma$ and the rows of $\mathbf{S}$ come from $(\mathbf{T}_{a,d}\boldsymbol{\alpha}_d)'$, where $d = 1, 2, \ldots, D$ serves to dimension index the $\mathbf{T}_a$ matrix and $\boldsymbol{\alpha}$ vector, with the first α parameter for each dimension constrained to 1 (e.g., $\alpha_{d,1} \equiv 1$). Both γ and $\boldsymbol{\alpha}_d$ are $(K-1) \times 1$ vectors of parameters. Thus, $\mathbf{S}$ can be partitioned as:

$$\mathbf{S} = \begin{bmatrix} (\mathbf{T}_{a,1}\boldsymbol{\alpha}_1)' \\ (\mathbf{T}_{a,2}\boldsymbol{\alpha}_2)' \\ \vdots \\ (\mathbf{T}_{a,D}\boldsymbol{\alpha}_D)' \end{bmatrix} \tag{4}$$

Both $\mathbf{T}_c$ and $\mathbf{T}_{a,d}$ are $K \times (K-1)$ matrices. Whereas Bock (1972) initially proposed the use of deviation contrasts common in the use of analysis of variance Thissen et al. (2010) and Thissen and Cai (in press) suggest the use of either a Fourier basis or identity-based $\mathbf{T}$ matrix (see also Thissen & Steinberg, 1986). For instance,

$$\mathbf{T_F} = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 1 & t_{22} & \ldots & t_2(K-1) \\ 2 & t_{32} & \ldots & t_3(K-1) \\ \vdots & \vdots & & \vdots \\ K-1 & 0 & \ldots & 0 \end{bmatrix} \tag{5}$$

has a linear first column, with the remaining columns determined by a Fourier basis with $t_{kp} = \sin[\pi(p-1)(k-1)/(K-1)]$. An identity-based matrix would be defined as:

$$\mathbf{T_I} = \begin{bmatrix} 0 & \mathbf{0}'_{K-2} \\ \mathbf{0}_{K-2} & \mathbf{I}_{K-2} \\ K-1 & \mathbf{0}'_{K-2} \end{bmatrix} \tag{6}$$

where $\mathbf{0}_{K-2}$ represents a $(K-2) \times 1$ vector of 0's and $\mathbf{I}_{K-2}$ is a $(K-2) \times (K-2)$ identity matrix. Use of either $\mathbf{T_F}$ or $\mathbf{T_I}$ and estimation of $\boldsymbol{\gamma}$ and for a particular dimension, $\mathbf{a}_d$ and $\alpha_{d,2}, \ldots, \alpha_{d,K-1}$, is equivalent to the least constrained model for that dimension where the scoring functions are fully estimated from the data. The use of $\mathbf{T_F}$ and $\mathbf{T_I}$ add additional utility when more constrained versions are used. For example, use of $\mathbf{T_F}$ while constraining $\alpha_{d,2}, \ldots, \alpha_{d,K-1}$ to 0 for a particular dimension is equivalent to a reparameterization of the GPC model. Fewer constraints on $\boldsymbol{\alpha}_d$ and the Fourier basis allow models that lie somewhere between the NRM and GPC, whereas additional constraints on slopes and/or intercepts may result in the partial credit (PC) and rating scale (RS) models (Thissen et al., 2010; Thissen & Cai, in press; Thissen & Steinberg, 1986). The GPC for a particular dimension may also be fit through the use of $\mathbf{T_I}$ by constraining $\alpha_{d,k} = k - 1$ for $\alpha_{d,2}, \ldots, \alpha_{d,K-1}$. The use of $\mathbf{T_I}$ also allows equality constraints on the values of the scoring functions (see Thissen et al., 2010; Thissen & Cai, in press).

The approach we use to fix scoring functions for estimation of response style factors is to place constraints on $\boldsymbol{\alpha}_d$ such that $\alpha_{d,1} = 1$ and $\alpha_{d,2}, \ldots, \alpha_{d,K-1}$ are 0, and then define the first column of $\mathbf{T}_{a,d}$ to reflect the desired scoring function values for that particular dimension. This is the same approach in the use of $\mathbf{T_F}$ in fitting the GPC model. Since all $\boldsymbol{\alpha}$ parameters are constrained in this approach, it may not be necessary to go through the process of reparameterizing the scoring function values as in the full MNRM. However, retaining this type of model within the framework of the MNRM allows flexibility, such as freely estimating the order of categories for one dimension, while constraining the scoring functions for another latent dimension. Researchers may also wish to compare models in which the scoring functions for a particular dimension are fixed versus freely estimated or vary between the GPC and NRM.

Appendix B. Complete Data Likelihood and Derivatives

We define the conditional mass function of Y_{ij} as:

$$f(y_{ij}|\mathbf{x}_i, \boldsymbol{\eta}_j) = \prod_{k=1}^{K_j} P(Y_{ij} = k|\mathbf{x}_i, \boldsymbol{\eta}_j)^{\chi_k(y_{ij})} \tag{7}$$

where $\chi_k(y_{ij})$ is an indicator function that equals 1 if $y_{ij} = k$ and 0 otherwise, $P(Y_{ij} = k|\mathbf{x}_i, \boldsymbol{\eta}_j)$ is the response function for item j and category k, and $\boldsymbol{\eta}_j = [\mathbf{a}_j', \boldsymbol{\alpha}_j', \boldsymbol{\gamma}_j']'$ contains all item parameters for item j. The conditional mass function of $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{in})'$ is:

$$f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\eta}) = \prod_{j=1}^{n} f(y_{ij}|\mathbf{x}_i, \boldsymbol{\eta}_j) \tag{8}$$

where $\boldsymbol{\eta}$ contains parameters across all items. For a single group, the complete data likelihood is:

$$L(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}) = \left[\prod_{i=1}^{N} f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\eta})\right] \left[\prod_{i=1}^{N} \phi(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]. \tag{9}$$

The log-likelihood allows separation of item parameters and parameters pertaining to the distribution of latent traits:

$$\log L(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Y}) = \log L(\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}) + \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) \tag{10}$$

The left-hand side results in the following, and resembles n independent multinomial logistic regression log-likelihoods:

$$l(\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}) = \log L(\boldsymbol{\eta}|\mathbf{X}, \mathbf{Y}) = \sum_{j=1}^{n} \left[\sum_{i=1}^{N} \sum_{k=0}^{K_j - 1} \chi_k(y_{ij}) \log P(Y_{ij} = k|\mathbf{x}_i, \boldsymbol{\eta}_j)\right] \tag{11}$$

If we assume that the latent traits are multivariate normal, this results in the following

for the right-hand side of (10):

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) = \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}) \propto \sum_{i=1}^{N} \left(-\frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) \tag{12}$$

While the derivatives for (12) are well-known, derivatives for item parameters can be obtained in the following manner. Dropping item and subject subscripts, the likelihood contribution of a single subject and item can be written as:

$$l = l(\boldsymbol{\eta}|\mathbf{x}, \mathbf{y}) = \sum_{k=0}^{K-1} \chi_k(y) \log \left[P(Y = k|\mathbf{x}, \mathbf{a}, \mathbf{S}, \mathbf{c}) \right]$$

The first order derivatives are given by the following:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} P_h (\mathbf{t}_{c,k} - \mathbf{t}_{c,h})' \right] \\
\frac{\partial l}{\partial \boldsymbol{\alpha}_d} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} P_h (\mathbf{S}_{d,k} - \mathbf{S}_{d,h})' \right] \mathbf{D}_{\mathbf{a}} \mathbf{x} \\
\frac{\partial l}{\partial \mathbf{a}} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} P_h (\mathbf{s}_k - \mathbf{s}_h) \right] \circ \mathbf{x}
\end{aligned}$$

where $P_h = P(Y = h|\mathbf{x}, \mathbf{a}, \mathbf{S}, \mathbf{c})$ is short-hand for the response function, $\mathbf{t}_{c,k}$ is the kth row of $\mathbf{T}_c$, $\mathbf{s}_k$ is defined as before (the kth column of $\mathbf{S}$), $\mathbf{D}_{\mathbf{a}} = \text{diag}(\mathbf{a})$ is a diagonal matrix, and $\mathbf{S}_{d,k}$ is defined as:

$$\frac{\partial \mathbf{s}_k}{\partial \boldsymbol{\alpha}_d} = \begin{bmatrix} \mathbf{0} \\ \vdots \\ \mathbf{t}_{a,d,k} \\ \vdots \\ \mathbf{0} \end{bmatrix} = \mathbf{S}_{d,k}$$

with $\mathbf{t}_{a,d,k}$ as the kth row of the $\mathbf{T}_a$ matrix for dimension d.

Then the second-order derivatives are given by:

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \boldsymbol{\gamma}} (\mathbf{t}_{c,k} - \mathbf{t}_{c,h}) \right] \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\alpha}_d'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \boldsymbol{\alpha}_d} (\mathbf{t}_{c,k} - \mathbf{t}_{c,h}) \right]' \\
\frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \mathbf{a}'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \mathbf{a}} (\mathbf{t}_{c,k} - \mathbf{t}_{c,h}) \right]' \\
\frac{\partial^2 l}{\partial \mathbf{a} \partial \mathbf{a}'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \mathbf{a}} (\mathbf{D}_{\mathbf{x}} (\mathbf{s}_k - \mathbf{s}_h))' \right] \\
\frac{\partial^2 l}{\partial \mathbf{a} \partial \boldsymbol{\alpha}_d'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \boldsymbol{\alpha}_d} (\mathbf{D}_{\mathbf{x}} (\mathbf{s}_k - \mathbf{s}_h))' + P_h (\mathbf{D}_{\mathbf{x}} (\mathbf{S}_{d,k} - \mathbf{S}_{d,h}))' \right]' \\
\frac{\partial^2 l}{\partial \boldsymbol{\alpha}_d \partial \boldsymbol{\alpha}_t'} &= \sum_{k=1}^{K} \chi_k(y) \left[\sum_{h=1}^{K} \frac{\partial P_h}{\partial \boldsymbol{\alpha}_t} (\mathbf{x}' \mathbf{D}_{\mathbf{a}} (\mathbf{S}_{d,k} - \mathbf{S}_{d,h})) \right]'
\end{aligned}$$

where $\mathbf{D}_{\mathbf{x}} = \text{diag}(\mathbf{x})$ is a diagonal matrix and:

$$\begin{aligned}
\frac{\partial P_h}{\partial \boldsymbol{\gamma}} &= P_h \left[\sum_{m=1}^{K} P_m (\mathbf{t}_{c,h} - \mathbf{t}_{c,m})' \right] \\
\frac{\partial P_h}{\partial \boldsymbol{\alpha}_d} &= P_h \left[\sum_{m=1}^{K} P_m (\mathbf{S}_{d,h} - \mathbf{S}_{d,m})' \right] \mathbf{D}_{\mathbf{a}} \mathbf{x} \\
\frac{\partial P_h}{\partial \mathbf{a}} &= P_h \left[\sum_{m=1}^{K} P_m (\mathbf{s}_h - \mathbf{s}_m) \right] \circ \mathbf{x}
\end{aligned}$$

Appendix C. Additional Estimation, Standard Error, and Scoring Details

The goal of MH-RM is to optimize the parameter estimates with respect to the observed data marginal log-likelihood:

$$l(\boldsymbol{\omega}|\mathbf{Y}) = \log L(\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}) = \sum_{i=1}^{N} \log \left[\int f(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\eta}) \phi(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x} \right] \tag{13}$$

Unfortunately, the D-fold integral in (13) makes computations intensive for numerical integration. MH-RM relies on the insight, due to Fisher's identity (Fisher, 1925), that:

$$\dot{l}(\boldsymbol{\omega}|\mathbf{Y}) = \int \dot{l}(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y}) \Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega}) d\mathbf{X} \tag{14}$$

We obtain draws from $\Pi(\mathbf{X}|\mathbf{Y}, \boldsymbol{\omega})$ to provide a Monte Carlo approximation to $\dot{l}(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$ using the same Metropolis-within-Gibbs sampler as in Cai (2010a). Proposal draws for the vector of latent traits are $\mathbf{x}_i^* = \mathbf{x}_i + \mathbf{e}_i$, with the increments drawn from a D-dimensional multivariate normal distribution, $\mathbf{e}_i \sim N(\mathbf{0}_D, c^2 \mathbf{I}_D)$, with tuning parameter c. Acceptance probabilities of the proposals can be computed by:

$$\min \left\{ \frac{f(\mathbf{y}_i|\boldsymbol{\eta}, \mathbf{x}_i^*) \phi(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{f(\mathbf{y}_i|\boldsymbol{\eta}, \mathbf{x}_i) \phi(\mathbf{x}_i^*|\boldsymbol{\mu}, \boldsymbol{\Sigma})}, 1 \right\} \tag{15}$$

The observed data gradient, can then be approximated by averaging across the complete data gradient of the m_p imputations,

$$\dot{l}(\boldsymbol{\omega}|\mathbf{Y}) \approx \mathbf{g}_{p+1} = \frac{1}{m_p} \sum_{t=1}^{m_p} \dot{l}(\boldsymbol{\omega}|\mathbf{X}_t^{(p+1)}, \mathbf{Y}) \tag{16}$$

And a recursive approximation to the conditional expectation of the complete data information matrix is computed by:

$$\boldsymbol{\Gamma}_{p+1} = \boldsymbol{\Gamma}_p + \beta_p \left\{ \frac{1}{m_p} \sum_{t=1}^{m_p} \mathbf{H}(\boldsymbol{\omega}^{(p)}|\mathbf{X}_t^{p+1}, \mathbf{Y}) - \boldsymbol{\Gamma}_p \right\} \tag{17}$$

where $\mathbf{H}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y}) = -\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'}$ is the complete data information matrix. The sequence of gain constants, $\{\beta_p; p \geq 1\}$, in the final stage of estimation meet the following conditions, ensuring that noise introduced by taking a small number of imputations at each iteration is filtered out:

$$\beta_p \in (0,1], \sum_{p=1}^{\infty} \beta_p = \infty, \text{ and } \sum_{p=1}^{\infty} \beta_p^2 < \infty$$

In the case of multiple groups, $g = 1, 2, \ldots, G$, the complete data log-likelihood is simply the sum of each group's log-likelihood, with group subscripts added to all item parameters, observed responses, and latent trait scores:

$$l(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y}) = \sum_{g=1}^{G} \left\{ l(\boldsymbol{\eta}_g|\mathbf{X}_g,\mathbf{Y}_g) + l(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g|\mathbf{X}_g) \right\} \tag{18}$$

The observed data gradient, $\dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})$, at iteration $p+1$ can be approximated by combining the average of complete data gradients for each of the groups,

$$\mathbf{g}_{p+1} = \begin{bmatrix} \mathbf{g}'_{p+1,1} & \mathbf{g}'_{p+1,2} & \cdots & \mathbf{g}'_{p+1,G} \end{bmatrix}' \tag{19}$$

where the second subscript provides the group index. The conditional expectation of the complete data information matrix is thus the super matrix:

$$\boldsymbol{\Gamma}_{p+1} = \begin{bmatrix} \boldsymbol{\Gamma}_{p+1,1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Gamma}_{p+1,2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Gamma}_{p+1,G} \end{bmatrix} \tag{20}$$

If there are any parameter constraints across groups, these may be implemented in via the strategy already presented in Section 4 and as also used by Cai (2010b), which may change (20) to something other than block-diagonal.

Standard errors under MH-RM are often obtained via the following equation due to (Louis, 1982), the elements of which can be approximated during or after estimation (see Cai, 2010a; Yang & Cai, 2014):

$$-\frac{\partial^2 l(\boldsymbol{\omega}|\mathbf{Y})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}'} = \int \left[\mathbf{H}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y}) - \dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})[\dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})]'\right]\Pi(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})d\mathbf{X} + \int \dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})\Pi(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})d\mathbf{X}\int [\dot{l}(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})]'\Pi(\mathbf{X}|\mathbf{Y},\boldsymbol{\omega})d\mathbf{X}$$

EAP scoring and marginal log-likelihood approximations (including transformations such as AIC and BIC) can be accomplished via Monte Carlo integration that entails averaging over Q draws for each $\mathbf{x}_i$ from $\phi(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$. For example, a marginal log-likelihood approximation of (13) is:

$$l(\boldsymbol{\omega}|\mathbf{Y}) \approx \sum_{i=1}^{N}\log\left[\frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\mathbf{x}_i^{(q)},\boldsymbol{\eta})\right] \tag{21}$$

Confidence intervals may be formed by approximating the variance of this quantity or transformations of it via the Delta method. For example, if we define $v_i = \frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\mathbf{x}_i^{(q)},\boldsymbol{\eta})$, then we can obtain the approximate variance of $-2l(\boldsymbol{\omega}|\mathbf{Y})$ by:

$$\text{var}\left(-2\sum_{i=1}^{N}\log\left[\frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\mathbf{x}_i^{(q)},\boldsymbol{\eta})\right]\right) \approx 4\sum_{i=1}^{N}\text{var}(\log(v_i)) \tag{22}$$

where

$$\text{var}(\log(v_i)) = \frac{1}{v_i^2}\text{var}\left(\frac{1}{Q}\sum_{q=1}^{Q} f(\mathbf{y}_i|\mathbf{x}_i^{(q)},\boldsymbol{\eta})\right) = \frac{1}{v_i^2(Q-1)}\sum_{q=1}^{Q}(f(\mathbf{y}_i|\mathbf{x}_i^{(q)},\boldsymbol{\eta}) - v_i)^2 \tag{23}$$

References

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response style in marketing research: A cross-national investigation. *Journal of Martketing Research*, *38*, 143–156.

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

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*, 179–197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*, 665-678.

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika*, *79*, 51–537.

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.

Bolt, D. M., Lu, Y., & Kim, J.-S. (in press). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*.

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*, 814-833.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*, 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307–335.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel item factor analysis and test

scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software Internatonal, Inc.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1-29.

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science*, *6*, 170-175.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*, 187-212.

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review*, *18*, 301-324.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, *60*, 151-174.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*, 104-115.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: method and application. In W. Gilks, S. Richardson, & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (p. 259-273). London: Chapman and Hall.

Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1-18.

Eddelbuettel, D., & Sanderson, C. (2014, March). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, *71*, 1054–1063.

Ferrando, P. J., Lorenzo-Seva, U., & Chico, E. (2003). Unrestricted factor analytic procedures for assessing acquiescent responding in balanced, theoretically unidimensionaly personality scales. *Multivariate Behavioral Research*, *38*, 353-374.

Fischer, R. (2004). Standardization to account for cross-cultural response bias: A classification of score adjustment procedures and review of research in JCCP. *Journal of Cross-Cultural Psychology*, *35*, 263-282.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700-725.

Fox, J. P. (2003). Stochastic EM for estimating the parameters of a multilevel IRT model. *British Journal of Mathematical and Statistical Psychology*, *56*, 65-81.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*, 176-188.

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the PROMIS smoking item banks. *Nicotine & Tobacco Research*, *16*, S175–S189.

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement*, *21*, 771-790.

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, *102*, 454-463.

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*, 116-138.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response styles. *Psychometrika*, *68*, 563-583.

Johnson, T. R., & Bolt, D. M. (2010). On the use of factor-analytic multinominal logit item response models to account for individual differences in response styles. *Journal of Educational and Behavioral Statistics*, *35*, 92-114.

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*, 161-177.

Kieruj, N. D., & Moors, G. (2010). Variations in response style behaviour by response scale format in attitude research. *International Journal of Public Opinion Research*, *22*, 320-342.

Kieruj, N. D., & Moors, G. (2013). Response style behavior: question format dependent or personal style? *Quality & Quantity*, *47*, 193-211.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213-236.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*, 201-228.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society - Series B*, *44*(2), 226-233.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics* (p. 105âĂŞ142). New York, NY: Academic Press.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*, 343-369.

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. sociodemographic correlates of gender role attitudes and perceptions of

ethnic discrimination reexamined. *Quality & Quantity*, *37*, 277-302.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities: A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, *20*, 303-320.

Morey, L. C. (1991). *The Personality Assessment Inventory: Professional manual*. Odessa, FL: Psychological Assessment Resources.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology*, *41*, 13-47.

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology: European Journal of Research Methods for the Beahvioral and Social Sciences*, *8*, 159-170.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measure of personality and social psychological attitudes.* San Diego, CA: Academic Press.

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

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response style. *Educational and Psychological Measurement*, *74*, 875–899.

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association*, *96*,

20-31.

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533–555.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393-408.

Thissen, D., & Cai, L. (in press). Nominal categories models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (2nd ed.). New York, NY: Chapman & Hall.

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* (p. 43-75). New York, NY: Taylor & Francis.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*, 522-547.

Trott, M. D., & Jackson, D. N. (1967). An experimental analysis of acquiescence. *Journal of Experimental Research in Personality*, *2*, 278–288.

van Rosmalen, J., van Herk, H., & Groenen, P. J. F. (2010). Identifying response styles: A latent-class bilinear multinominal logit model. *Journal of Marketing Research*, *46*, 157-172.

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement*, *43*, 335-353.

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model.

Journal of Educational Measurement, *48*, 441-456.

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178-189.

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monro algorithm. *Journal of Educational and Behavioral Statistics*, *39*, 550–549.

Table 1: Item parameters from Item 1 of Hedonic Benefits under 4 different response style models.

Parameter	ERS	MRS	EMRS	ERS and MRS
a_1	1.48	1.47	1.49	1.48
a_2	1.74	0.83	0.93	1.69
a_3				0.75
c_1	0.00	0.00	0.00	0.00
c_2	2.22	2.01	2.20	2.21
c_3	3.07	2.70	2.84	3.03
c_4	2.16	1.89	2.06	2.11
c_5	-0.18	0.21	-0.04	-0.24

Note. a_1 always refers to the slope on hedonic benefits and a_2 to the slope of the response style factor. In the case of ERS and MRS, a_2 is for ERS and a_3 is for MRS. For identification, estimation of intercept parameters is done via estimating γ, i.e., $\mathbf{c} = \mathbf{T}_c\gamma$ (see also Appendix A).

Table 2: Possible scoring functions for a hypothetical 5-category graded item.

Scoring Function	Latent trait
$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	Substantive trait
$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \end{bmatrix}$	Extreme Response Style (ERS)
$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$	Midpoint Response Style (MRS)
$\begin{bmatrix} 2 & 1 & 0 & 1 & 2 \end{bmatrix}$	Hybrid Extreme-Mid Response Style (EMRS)
$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 \end{bmatrix}$	Acquiescence bias (slopes equal across items)
$\begin{bmatrix} 0 & 0 & 0 & 1 & 1 \end{bmatrix}$	Acquiescence as a tendency to respond above the midpoint
$\begin{bmatrix} 1 & 0 & 1 & 0 & 1 \end{bmatrix}$	Use of extreme and midpoint anchors
$\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$	Socially Desirable Responding*
$\begin{bmatrix} 0.8 & 1.6 & 2.8 & 2.9 & 2.2 \end{bmatrix}$	Socially Desirable Responding with varying weights**

Note. * Assuming the fourth category is the most socially desirable response. ** Non-integer weights representing degree of social desirability for each category based on data (e.g., "Spontaneous, impulsive" at five trait levels in a "general" context; Kuncel & Tellegen, 2009).

Table 3: Constraints for common and fitted models.

Model	Slope	SF	SF Constraints	Intercept for category k
Full nominal (NRM-free)	a_{jd}	$\mathbf{s}_{jd}$	Estimated	c_{jk}
Stereotype (ST)	a_{jd}	$\mathbf{s}_{d}$	Estimated	$c_j + \beta_k$
Generalized partial credit (GPC)	a_{jd}	$\mathbf{s}_{jd}$	Fixed to $[0 \quad 1 \quad \ldots \quad K_j - 1]$	c_{jk}
Partial credit (PC)	a_{d}	$\mathbf{s}_{jd}$	Fixed to $[0 \quad 1 \quad \ldots \quad K_j - 1]$	c_{jk}
Rating scale (RTS)	a_{d}	$\mathbf{s}_{jd}$	Fixed to $[0 \quad 1 \quad \ldots \quad K_j - 1]$	$c_j + \beta_k$
Response style	a_{jd}	$\mathbf{s}_{jd}$	Fixed to user-specified values	c_{jk}
NRM-equal	a_{d}	$\mathbf{s}_{d}$	Estimated	c_{jk}
NRM-fixed	a_{d}	$\mathbf{s}_{jd}$	Fixed to user-specified values	c_{jk}

Note. SF = Scoring function. Item slopes and scoring functions are particular for a given dimension, d, with $\mathbf{s}_{jd}$ indicating row d in $\mathbf{S}$ for item j. Omitted j subscripts indicate equality across items. In the case of GPC, PC, RTS, and NRM-fixed models, j varies for scoring functions because of possible differences in the number of categories across items. Parameterizations assume fixed variance of latent traits (in the case of single-group models). Alternative parameterizations of PC, RS, and NRM-fixed models could include fixing item slopes (e.g., to 1) and freeing the variance of the latent trait. Appendix A describes how scoring functions and intercepts are re-parameterized as $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively, to allow for identification and estimation.

Table 4: Models fit to PROMIS® smoking data.

Model	Dimensions	**a**	# param.	AIC	BIC
1	6D	free	732	[768271, 768394]	[773096, 773218]
2	6D, ERS	free, free	884	[744698, 744862]	[750525, 750688]
3	6D, ERS	free, equal	747	[746216, 746376]	[751140, 751300]
4	6D, MRS	free, free	884	[762042, 762192]	[767869, 768018]
5	6D, MRS	free, equal	747	[762118, 762267]	[767042, 767191]
6	6D, EMRS	free, free	884	[750358, 750515]	[756184, 756342]
7	6D, EMRS	free, equal	747	[751288, 751445]	[756212, 756369]
8	6D, ERS, MRS	free, free free	1038	[740394, 740586]	[747236, 747427]
9	6D, ERS, MRS	free, equal, free	901	[741941, 742130]	[747880, 748068]
10	6D, ERS, MRS	free, free, equal	901	[740418, 740609]	[746356, 746547]
11	6D, ERS, MRS	free, equal, equal	764	[742159, 742348]	[747195, 747384]
12	6D, ERS	equal, equal	615	[755334, 755483]	[759387, 759536]
13	6D, MRS	equal, equal	615	[772884, 773018]	[776937, 777071]
14	6D, EMRS	equal, equal	615	[760698, 760842]	[764752, 764896]
15	6D, ERS, MRS	equal, equal, equal	632	[751107, 751282]	[755272, 755448]
16	6D, NRM	free, equal	750	[745423, 745587]	[750366, 750530]
17	6D ST, NRM	free, equal	357	[758617, 758776]	[760970, 761129]

Note. "**a**" indicates whether slopes were free across items or constrained equal across items for each dimension listed. See Table 2 for example scoring functions for ERS, MRS, and EMRS. AIC and BIC are 95% CI estimates based on a Monte Carlo integration estimate of the marginal log-likelihood for each model.

Table 5: Estimated latent variable means and variance-covariance matrices for Model 10.

	HB	ND	HR	PSR	CB	SB	ERS	MRS
Daily smokers								
HB	1.00							
ND	0.52	1.00						
HR	0.06	0.46	1.00					
PSR	0.14	0.52	0.83	1.00				
CB	0.71	0.72	0.32	0.39	1.00			
SB	0.72	0.67	0.30	0.34	0.77	1.00		
ERS	0.21	0.09	0.03	0.05	0.10	0.14	1.00	
MRS	0.00	-0.04	-0.00	-0.03	-0.02	-0.01	-0.12	1.00
Mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Non-daily smokers								
HB	1.04							
ND	0.66	1.44						
HR	0.22	0.73	1.09					
PSR	0.17	0.64	0.78	0.98				
CB	0.59	0.86	0.43	0.40	1.00			
SB	0.82	0.70	0.42	0.34	0.69	1.24		
ERS	0.16	0.03	-0.02	-0.04	-0.01	0.04	0.96	
MRS	0.03	0.04	-0.03	-0.04	-0.01	-0.05	-0.04	1.04
Mean	-0.38	-1.12	-0.22	-0.11	-0.55	-0.32	-0.04	-0.17

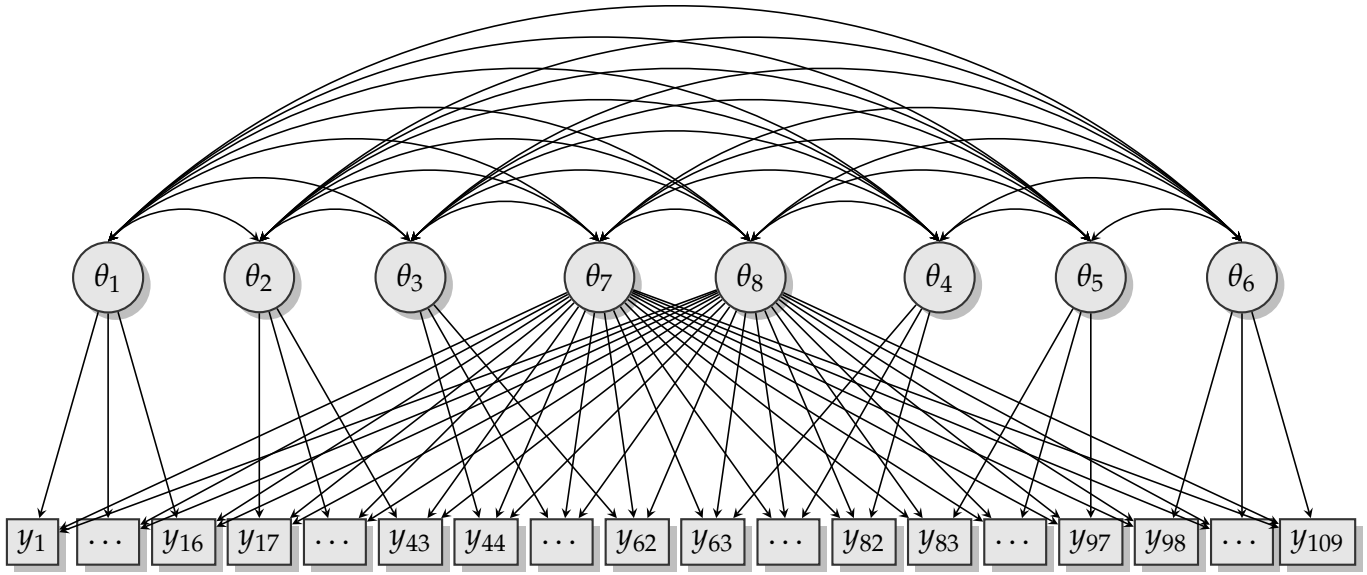
Note. HB = Hedonic benefits; ND = Nicotine dependence; HR = Health risks; PSR = Psychosocial risks; CB = Coping benefits; SB = Social benefits; ERS = Extreme response style; MRS = Midpoint response style. Means and variances are fixed to 0 and 1, respectively, for daily smokers.

Table 6: Latent trait recovery for Simulation 1.

	GPC		RS		RS-cor	
	r	RMSE	*r*	RMSE	*r*	RMSE
GPC Data Generating Model						
Hedonic benefits	0.97	24.48	0.97	24.52	0.97	24.52
Nicotine dependence	0.98	18.24	0.98	18.30	0.98	18.29
RS Data Generating Model						
Hedonic benefits	0.92	39.84	0.96	29.65	0.96	29.65
Nicotine dependence	0.94	35.50	0.97	23.69	0.97	23.69
Extreme response style			0.91	41.49	0.91	41.50
Midpoint response style			0.77	63.80	0.77	63.84
RS-cor Data Generating Model						
Hedonic benefits	0.91	41.14	0.95	29.97	0.96	29.62
Nicotine dependence	0.93	36.41	0.97	23.92	0.97	23.73
Extreme response style			0.91	41.79	0.91	40.74
Midpoint response style			0.77	64.34	0.78	62.85

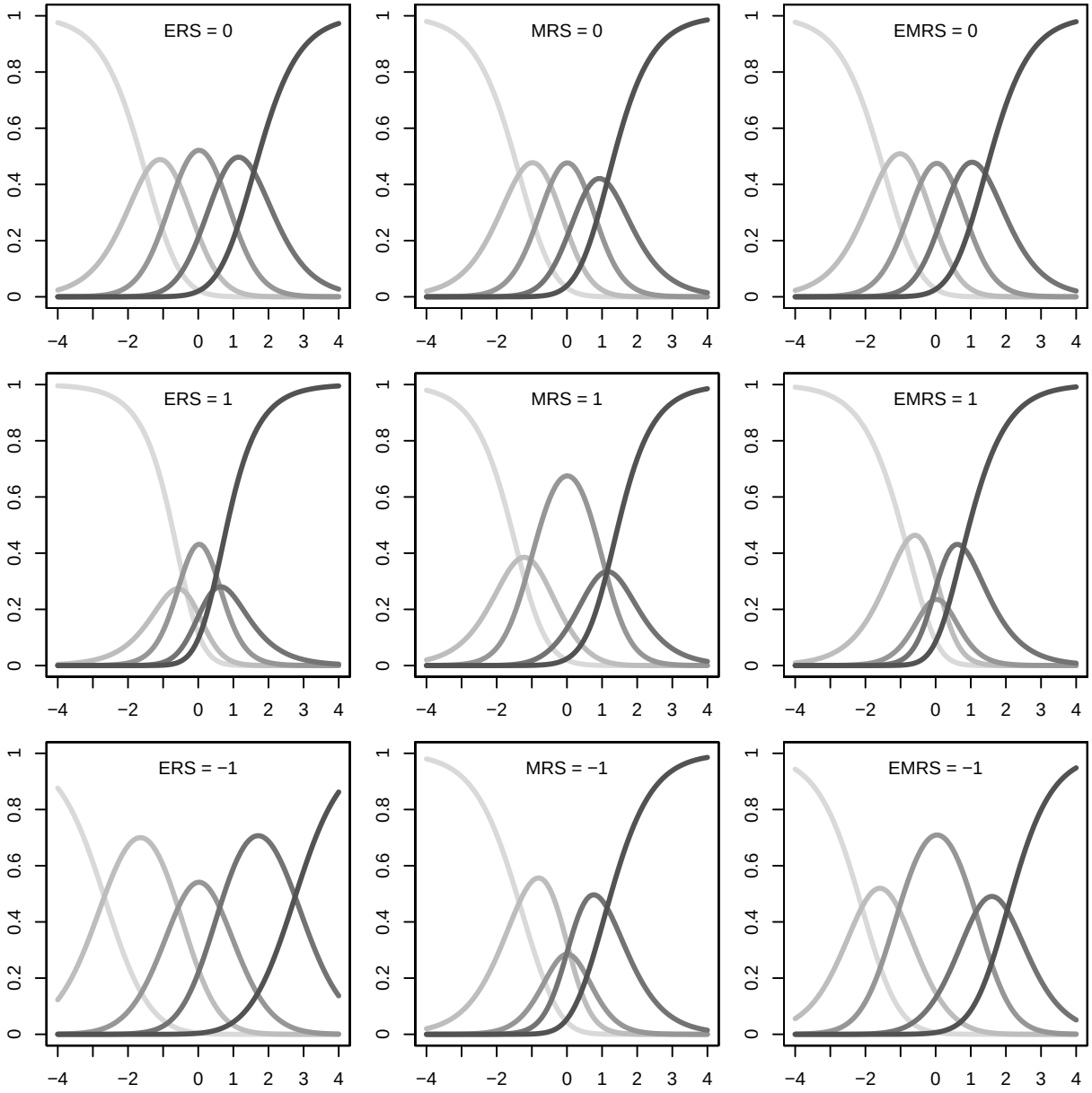
Note. Columns correspond to each fitted model. GPC = Generalized partial credit model with only substantive traits; RS = Model with additional response styles (ERS and MRS) that are uncorrelated, RS-cor = Model with additional response styles (ERS and MRS) that are correlated with each other and substantive traits.

Figure 1: Six-dimensional model plus two additional response style factors.



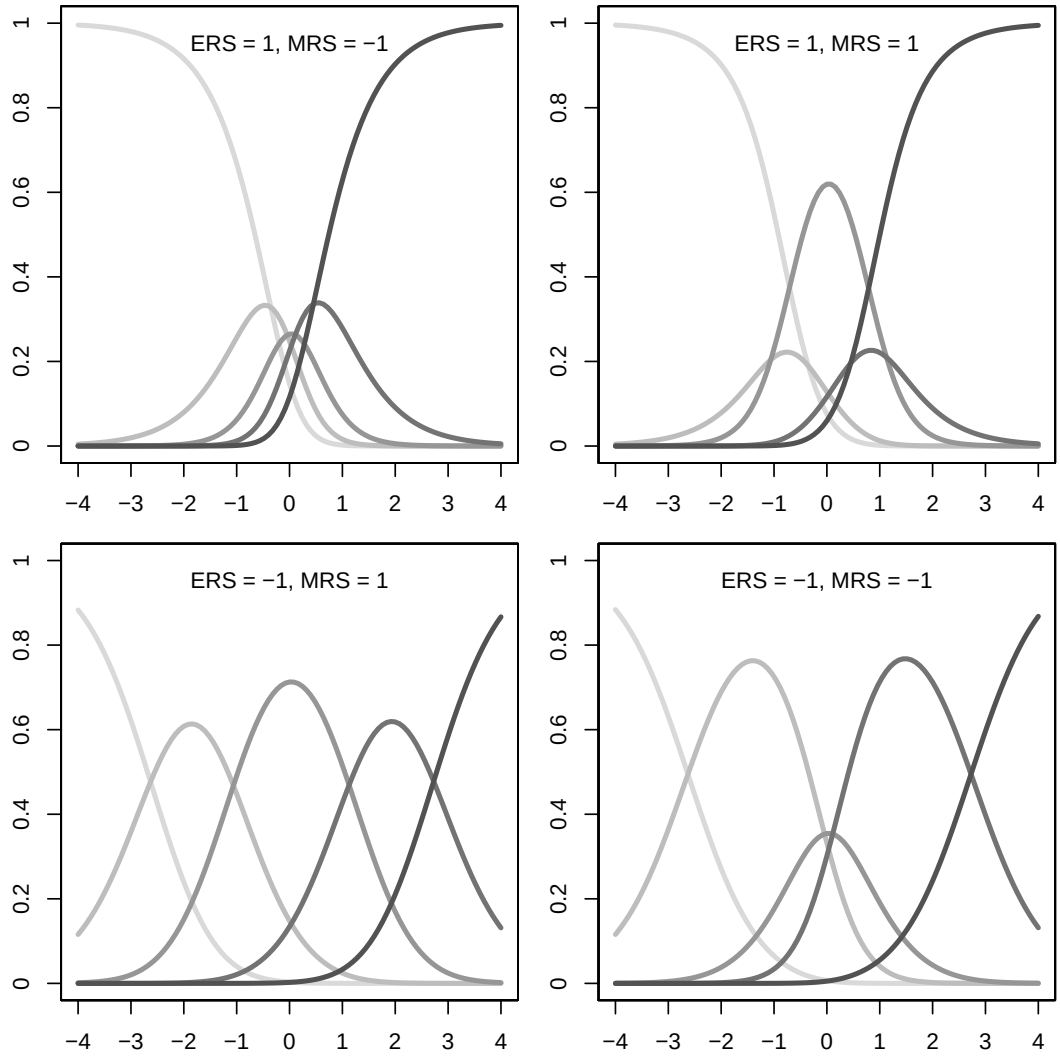
Note. θ_1 = Hedonic benefits; θ_2 = Nicotine dependence; θ_3 = Health risks; θ_4 = Psychosocial risks; θ_5 = Coping benefits; θ_6 = Social benefits; θ_7 and θ_8 are two hypothetical response style dimensions.

Figure 2: Example item response functions for a hedonic benefits item from three models that each include a single response style factor.



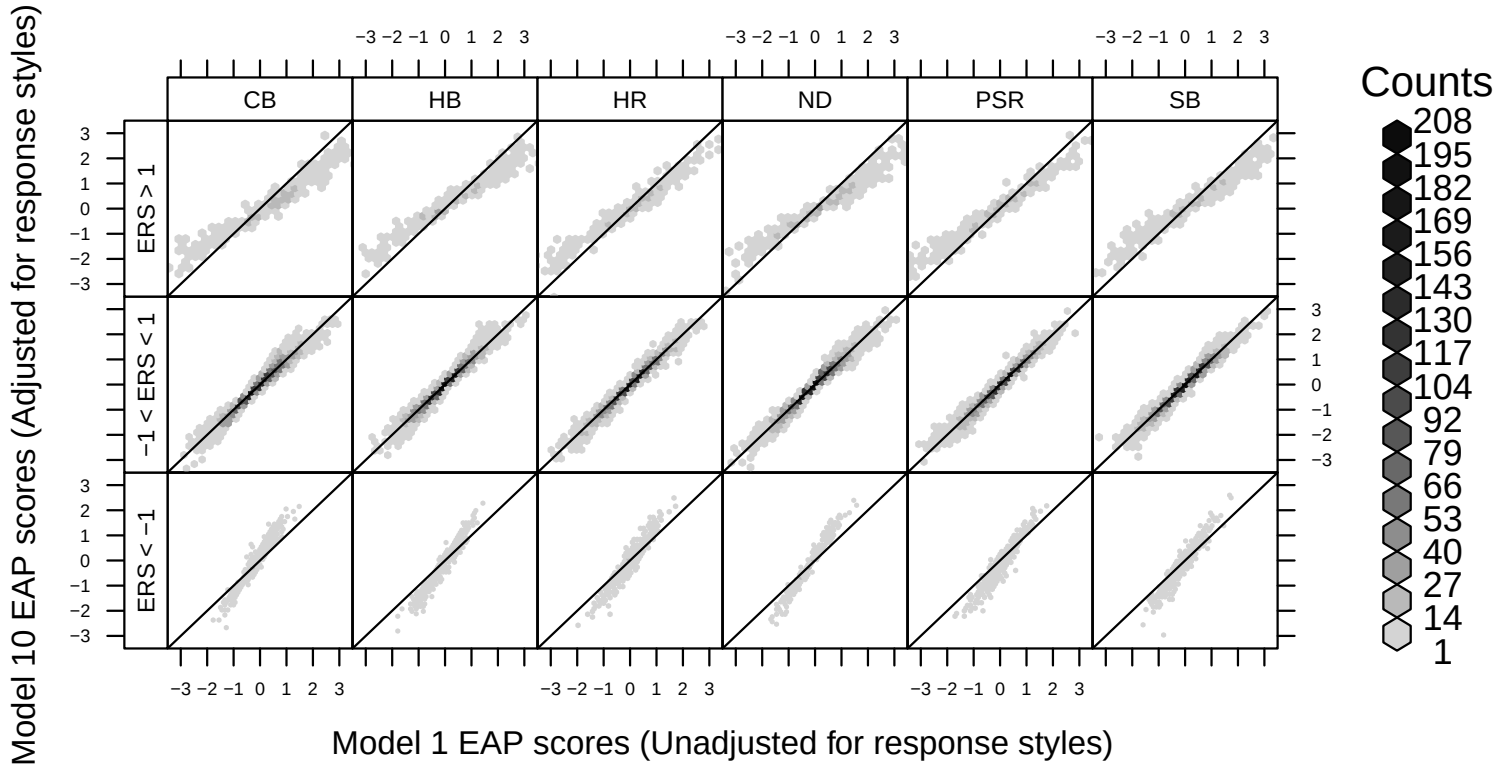
Note. Each column contains a cross-section of item response functions from a model at three levels of that model's response style factor. Top row is at the mean of the response style factor (0), middle row is high on the response style (1) and bottom row is low on the response style (-1). ERS = Extreme response style $\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \end{bmatrix}$, MRS = Midpoint response style $\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$, EMRS = Hybrid of extreme and midpoint response style at opposite poles the same dimension $\begin{bmatrix} 2 & 1 & 0 & 1 & 2 \end{bmatrix}$.

Figure 3: Example item response functions for a hedonic benefits item from a single model that includes two response style factors.



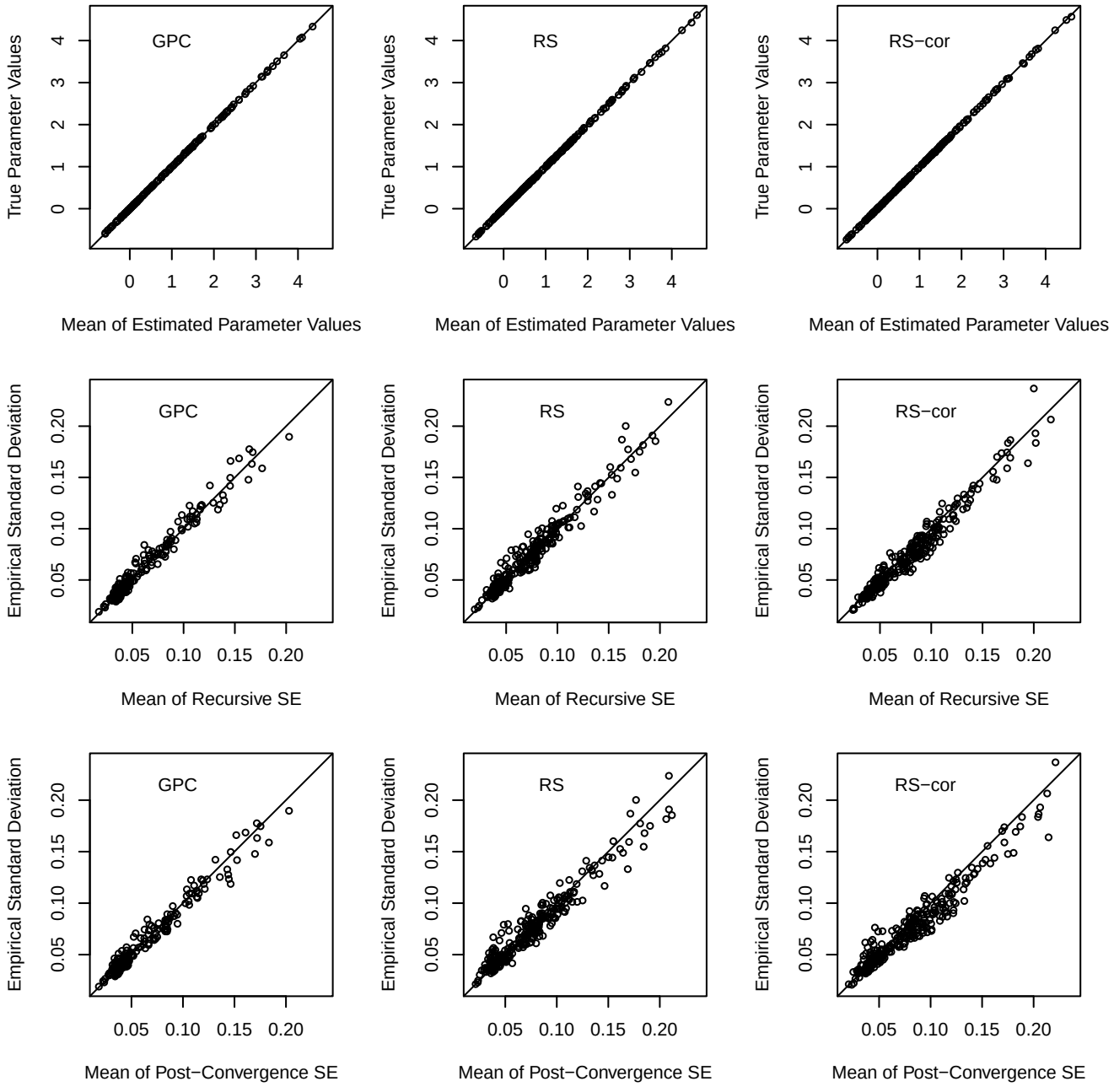
Note. Response functions are cross-sections from a single model that contains separate extreme response style (ERS) $\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \end{bmatrix}$ and midpoint response style (MRS) $\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$ factors.

Figure 4: Substantive trait EAP scores at three grouped levels of ERS.



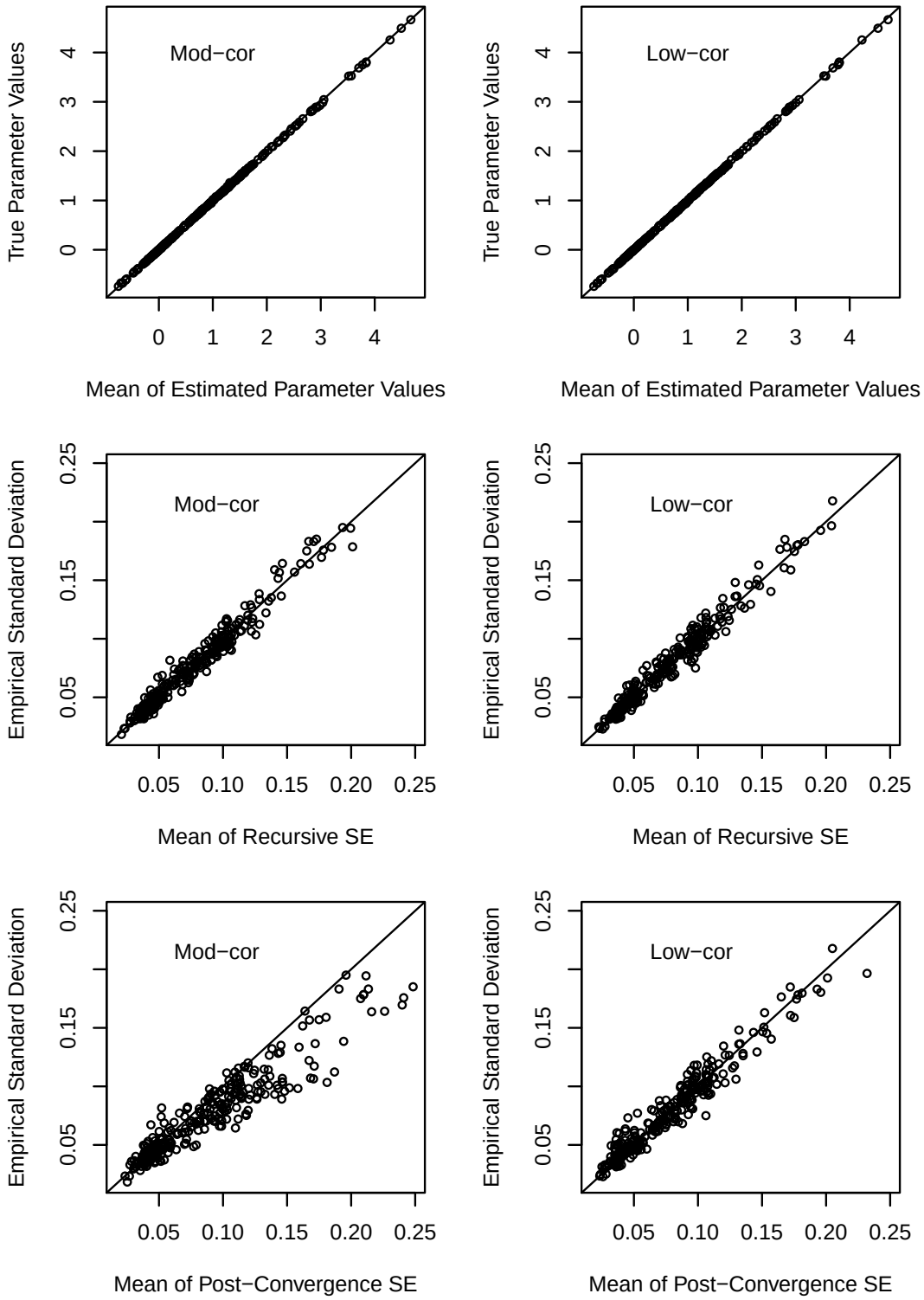
Note. CB = Coping benefits; HB = Hedonic benefits; HR = Health risks; ND = Nicotine dependence; PSR = Psychosocial risks; SB = Social benefits.

Figure 5: Simulation 1 parameter recovery and standard error calibration.



Note. Each column is from a different data generating model, from left to right: Generalized partial credit model with substantive traits only (GPC), the GPC model with the addition of orthogonal extreme response style and midpoint response style factors (RS), and extreme response style and midpoint response style factors that are correlated with each other and the substantive dimensions (RS-cor). Each row examines a different outcome, from top to bottom: Parameter recovery, recursive standard errors, and post-convergence standard errors. SE = Standard errors.

Figure 6: Simulation 2 parameter recovery and standard error calibration.



Note. Each column is from a different data generating model, from left to right: ERS and ARS with a moderate correlation (Mod-cor) of .46 between two substantive factors, and ERS and ARS with a low correlation (Low-cor) of .05 between two substantive factors. Each row examines a different outcome, from top to bottom: Parameter recovery, recursive standard errors, and post-convergence standard errors. SE = Standard errors.