Modeling Response Style Using Vignettes and Person-Specific Item Response Theory

Katherine G. Jonas and Kristian E. Markon

Department of Psychological and Brain Sciences

University of Iowa

Author Note

Correspondence may be addressed to Katherine G. Jonas, Stony Brook Health Sciences Center T10-060, 101 Nicholls Road, Stony Brook NY, 11794; katherine.jonas@stonybrookmedicine.edu

Abstract

Responses to survey data are determined not only by item characteristics and respondents' trait standings, but also by response styles. Recently, methods for modeling response style with personality and attitudinal data have turned toward the use of anchoring vignettes, which provide fixed rating targets. While existing research is promising, a few outstanding questions remain. First, it is not known how many vignettes and vignette ratings are necessary to identify response style parameters. Second, the comparative accuracy of these models is largely unexplored. Third, it remains unclear whether correcting for response style improves criterion validity. Both simulated data and data observed from a population-representative sample responding to a measure of personality pathology (the Personality Inventory for DSM-5; PID-5) were modeled using an array of response style models. In simulations, most models estimating response styles outperformed the GRM, and in observed data, all response style models were superior to the GRM. Correcting for response style had a small, but in some cases significant, effect on the prediction of self-reported social dysfunction.

Keywords: response style, item response theory, vignettes, personality assessment

Modeling Response Style Using Vignettes and Person-Specific Item Response Theory

Introduction

Response Styles

Individuals respond to survey items in certain ways regardless of item content, termed response styles. Common response styles include a disacquiescent response style (DRS), in which a person preferentially uses the lower range of the response scale; acquiescent response style (ARS), in which they tend to use the upper range of the scale; the central or mid-point response style (MRS), in which a person prefers the center of the scale; and extreme response style (ERS), where a person tends to use ends of the response scales.

Error introduced by response styles can cause an array of unintended effects. Response styles can artificially inflate scale correlations and estimates of longitudinal trait stability (Baumgartner & Steenkamp, 2001; Weijters, Geuens, & Schillewaert, 2010). Shared response styles increase estimates of inter-rater agreement (Dolnicar & Grün, 2009), and decrease them if response styles differ between raters. If response styles covary with group identity—nationality, for example—response styles manifest as differential item functioning (Bolt & Johnson, 2009; Herk, Poortinga, & Verhallen, 2004) and false between-group differences (Grol-Prokopczyk, Freese, & Hauser, 2011; Mõttus et al., 2012).

And yet, whether and how response styles should be modeled is an unresolved issue. Models incorporating response style do not always improve upon the fit of standard IRT models (Wetzel, Böhnke, & Rose, 2016), nor do they necessarily result in significant changes to trait estimates (Plieninger, 2017). This is especially true when response style is correlated with the trait of interest, as is often the case in personality research (He, Bartram, Inceoglu, & Van de Vijver, 2014). As a result, trait estimates corrected for response style often have poorer criterion validity than uncorrected estimates (Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; McCrae et al., 1989; Wetzel et al., 2016; McGrath, Mitchell, Kim, & Hough, 2010).

However, recent analyses have investigated new, promising methods of modeling

response style. Ferrando (2014) found response style moderated the criterion validity of a measure of test anxiety. Bolt, Lu, and Kim (2014) showed that correcting countries' mean level of conscientiousness for response style changed their rank-ordering. Wetzel and Carstensen (2015) observed that incorporating estimates of ERS and DRS improved model fit. In short, further research is needed to determine the conditions under which modeling response style is worthwhile, in terms of model fit and criterion validity.

Modeling Response Style

Latent variable and item-response theory models. Perhaps because response styles are described categorically (as either ARS, DRS, MRS or ERS), latent class models are often used to represent them. In these models, a prototypical response style defines the latent class, and class membership probabilities reflect individuals' response style (Morren, Gelissen, & Vermunt, 2011; Moors, 2012). Class mixture models function similarly, with the exception that latent classes are derived empirically (Bolt, Cohen, & Wollack, 2001). The limitation of these models is that in the LCA, response styles are fixed by the prototype, which may not represent many individuals. Both models assume response styles are categorical, an assumption which is likely not met in reality. Latent trait models address this second weakness by treating response style as a continuous trait. In some models, this trait is still defined by a prototype response style (e.g. ERS and MRS traits Bolt & Johnson, 2009; Falk & Cai, 2016). However, this is not strictly required.

Indeed, three recent studies have developed models that treat response style as a free parameter estimated for each individual (Bolt et al., 2014; Jin & Wang, 2014; Ferrando, 2014). We call these methods person-specific item response theory (PS IRT) models, because they are most easily conceptualized as models that estimate difficulty and discrimination parameters—items parameters in IRT models—for each individual (it is worth noting, however, that any latent trait model could be reparameterized as a PS IRT model). Bolt et al. (2014) assessed response style by estimating a vector of item category intercepts for each individual, and found this to be a significant improvement over a simple

nominal response model.[1] Similarly, Jin and Wang (2014) supplement the traditional partial credit model a person-specific weighting parameter, ω_i that acts as a moderator of item category thresholds (this model is akin to the latent trait model developed by Rossi, Gilula, and Allenby (2001) translated into an IRT framework). Ferrando (2014) focuses on individual differences in discrimination, estimating an a_i parameter to capture what might be described as attentiveness in achievement or opinion tests, or traitedness in personality tests. Ferrando found a_i moderated the magnitude of the correlation between test anxiety and scores on a final statistics exam.

Using vignette ratings. The model developed by Bolt et al. (2014) is unique in that it incorporates vignettes into a PS IRT model. Vignettes distinguish response style variance from trait variance, because if a vignette is assumed to have a true trait score, differences in ratings should reflect response style. Those response style estimates can then be used to adjust trait estimates from self-report data (King, Murray, Salomon, & Tandon, 2004). Without vignettes, researchers must constrain response styles traits to be orthogonal to content traits, in order to distinguish substantive and response style variance. However, it is known that content traits and response style traits are, in fact, correlated (He et al., 2014). Vignettes are useful because they are objective rating targets that can be used to identify response style traits, which can then be allowed to correlate with content traits. Using vignettes to identify response styles, Bolt et al. (2014) found that the updated estimates were sufficiently different from a traditional IRT model to change the rank order of countries' mean levels of Conscientiousness. In this approach, however, vignettes were assumed to reflect *only* response style estimates, preventing the distinction of vignette severity and ARS and DRS.

The validity of the use of vignettes depends on a number of assumptions (King et al.,

[1] Johnson (2003) achieves a similar end by estimating individual variability around category thresholds as a random effect. Johnson (2003) shows that neglecting response style has significant adverse effects on inference. However, since this method does not provide an interpretable quantification of response style, it is not discussed further here.

2004). First, it is assumed that all respondents perceive the vignette as reflecting the same trait. Second, the trait being assessed in the vignette is the same as that being assessed in self-report measures. Third, items are assumed to function in the same way in both self and vignette rating contexts. These assumptions allow item parameters to be constrained to be equal across vignette and self ratings, and between people. Tests of differential item functioning can be used to test some of these assumptions empirically.

The present study

The models tested by Bolt et al. (2014), Ferrando (2014), Jin and Wang (2014), among others, mark significant steps forward in modeling response styles, and vignettes allow for a relatively pure measure of scale usage. There are, however, a number of outstanding issues. First, it is unclear how many vignettes and vignette ratings are necessary to estimate response style parameters accurately. Second, few models have been evaluated in terms of their ability to recover simulated parameters, and few have been directly compared to one another (a notable exception being Bolt et al., 2014). Third, the evidence on whether response style models improve criterion validity remains mixed, especially for outcomes related to personality and psychopathology. It may be that the utility of modeling response styles changes substantially when applied to vignette data.

Aim one: How many vignettes is enough? It remains unknown how many vignettes and how many ratings per vignette are necessary to accurately estimate response styles. In King et al. (2004), it is recommended that two to three vignettes be rated for each self-report trait. By contrast, Bolt et al. (2014) collected 30 vignette responses for a single trait. Generally, recommendations have been based on common sense, rather than empirically. We simulate tests in which the number of vignettes varies, as does the number of items used to rate each vignette, and assess recovery of response style parameters through Kullback-Liebler divergence (KL divergence), a measure of "Bayesian learning" Xie and Carlin (2006). KL divergence quantifies the difference between prior and posterior distributions, and reflects the amount of information learned from the data (Kullback &

Leibler, 1951):

$$D_{KL}\Big(p(\theta|Y)|p(\theta)\Big) = \int p(\theta|Y) \log \left(\frac{p(\theta|Y)}{p(\theta)} \right) d\theta \quad (1)$$

Where $p(\theta|Y)$ is the posterior distribution of a parameter θ given observed data Y, and $p(\theta)$ is the prior distribution. When the data is not at all informative with regard to a parameter, the numerator and denominator are equal, and divergence is 0. Larger values reflect more information has been gained.

Aim two: Model accuracy & comparison. Our second aim was to assess and compare the accuracy of the models that have been proposed. Specifically, we compare models of ERS and MRS (Jin & Wang, 2014), ARS and DRS (similar to those proposed by King et al., 2004 and Bolt et al., 2014, adapted to the graded response model), inattention (Ferrando, 2014), and combinations thereof.[2] The models are briefly described below.

GRM. The traditional graded response model defines the probability that individual i responds to item j by endorsing response option k or higher by the equation (Samejima, 1969):

$$P(Y_{ij} \geq k) = \frac{exp(a_j\theta_i - b_{jk})}{1 + exp(a_j\theta_i - b_{jk})} \quad (2)$$

Where θ_i is the latent trait of individual i, a_j is the discrimination of item j, and b_{jk} is the response category threshold k for item j. The probability that an individual endorses category k is $P(Y_{ij} \geq k) - P(Y_{ij} \geq k + 1)$. Since we apply the GRM to a multi-trait personality assessment, the θ_i parameter is in fact a vector of traits $\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2} \ldots \theta_{i5})$. However, the test has a simple structure (i.e., each item measures a single trait), so the

[2] We do not include Bolt's exact model because it is based on the nominal response model, while others are based on the graded response model, preventing direct comparisons. We attempted to estimate a version of Bolt's model that included k-1 b_{ik} parameters for each individual, which could account for ERS, MRS, ARS, and DRS, and other idiosyncratic response styles. Convergence was generally poor, because certain combinations of b_{ik} values can result in un-ordered category thresholds, violating the category ordering constraints of the GRM. Developments in the software used to estimate these models may soon make this model feasible.

model is unidimensional at the item level. Vector and matrix notation are therefore suppressed for reasons of clarity.

b_i model. The b_i model includes a single person-specific difficulty, b_i, for each respondent:

$$P(Y_{ij} \geq k) = \frac{exp(a_j\theta_i - (b_{jk} + b_i))}{1 + exp(a_j\theta_i - (b_{jk} + b_i))} \tag{3}$$

The b_i parameter reflects ARS and DRS by shifting the category response function by a constant. Figure 1b and 1c show how the b_i parameter shifts category response curves (relative to the GRM, depicted in Figure 1a). When b_i is negative (ARS), curves are shifted downward along the latent trait, and the probability of using a higher response category for a given θ_i increases (ARS). When b_i is positive (Figure 1c), the curves shift to the right, and the probability of using a lower response option (DRS) increases.

a_i model. The a_i model incorporates a person-specific discrimination parameter (Ferrando, 2014):

$$P(Y_{ij} \geq k) = \frac{exp(a_j a_i\theta_i - b_{jk})}{1 + exp(a_j a_i\theta_i - b_{jk})} \tag{4}$$

The a_i parameter allows for individual differences in the strength of the relationship between item responses and θ_i. Figures 2b and 2c show the effect of the a_i parameter on category response curves, relative to the GRM (2a). When a_i is less than one (2b), the category response function is flattened, meaning that the probability of endorsing any given category becomes more random, reflecting inattention or lack of traitedness. When a_i is greater than one (2c), category response functions become more steep, increasing the perceived difference between two response options, reflecting attention or traitedness.

$a_i + b_i$ model. The a_i model does not allow for the estimation of ARS or DRS. To encompass those tendencies, we combine the a_i and b_i models:

$$P(Y_{ij} \geq k) = \frac{exp(a_j a_i\theta_i - (b_{jk} + b_i))}{1 + exp(a_j a_i\theta_i - (b_{jk} + b_i))} \tag{5}$$

The resulting model reflects attentiveness/traitedness, as well as ARS/DRS.

ω_i model. This model includes a person-specific category threshold weighting parameter ω_i (Jin & Wang, 2014):

$$P(Y_{ij} \geq k) = \frac{exp(a_j\theta_i - \omega_i b_{jk})}{1 + exp(a_j\theta_i - \omega_i b_{jk})} \tag{6}$$

ω_i moderates the distance between category thresholds. This effect is depicted in Figure 3. When ω_i less than one (ERS, Figure 3b), the distance between category thresholds narrows, making it more likely that the respondent will endorse the lowest or highest category. When ω_i is greater than one (MRS; Figure 3c), the interval between category thresholds widens, and the respondent is more likely to use one of the middle response options.

$\omega_i + b_i$ model. Like the a_i model, the ω_i model does not capture ARS and DRS. We combine it with the b_i model to do so:

$$P(Y_{ij} \geq k) = \frac{exp(a_j\theta_i - \omega_i(b_{jk} + b_i))}{1 + exp(a_j\theta_i - \omega_i(b_{jk} + b_i))} \tag{7}$$

This results in a model encompassing MRS, ERS, ARS and DRS.

Equations 2 through 7 model self-report data. In vignette ratings, the θ_i parameter is replaced with the vignette's trait standing, v. For example, in the b_i model, the probability of respondent i rating vignette v in category k or greater on item j is:

$$P(Y_{ivj} \geq k) = \frac{exp(a_j v - (b_{jk} + b_i))}{1 + exp(a_j v - (b_{jk} + b_i))} \tag{8}$$

In sum, the rating is a function of the item parameters, the vignette severity, v, and the individual's response style, b_i. Multiple items are used to rate each vignette, and item parameters a_j and b_{jk} are fixed across self- and vignette-report items. This structure allows item parameters to be distinguished from the vignette severity, v. An individual's ARS or DRS is assumed to be constant across traits. However, one could estimate b_i as a vector of parameters, assuming the availability of vignette ratings for each trait.

To compare these models, we simulated date from each of them, fit both the data-generating model and the GRM to simulated data, and compared them in terms of model fit and the precision and accuracy of parameter estimation. Since a number of

studies have shown the correlation between response style traits and content traits to be particularly important (Bolt & Johnson, 2009; Grol-Prokopczyk et al., 2011; Mõttus et al., 2012), we also report a set of simulations in which the correlation between b_i and θ_i is systematically varied, in order to test the effect of this correlation on model performance.

Aim three: Criterion validity. A third limitation of the existing literature is that very few studies have measured the effect of response style on criterion validity. To our knowledge, only Ferrando (2014) and Rossi et al. (2001) have done so, despite improving criterion validity being the basic motivation for identifying response styles. With these concerns in mind, we tested the effect of modeling response style on the relationship between self-reported personality and functioning in work, leisure, and family roles.

Methods

Observed Data.

The Personality Inventory for DSM-5 (PID-5) is a self-report measure of personality psychopathology (Krueger, Derringer, Markon, Watson, & Skodol, 2012). The 220 items of the PID-5 comprise five major traits (Negative Affect, Detachment, Antagonism, Disinhibition, and Psychoticism), which can be parsed into 25 specific traits. Five sub-traits were used in the current analyses: Emotional Lability (from the Negative Affectivity domain), Withdrawal (Detachment), Manipulativeness (Antagonism), Irresponsibility (Disinhibition), and Unusual Beliefs and Experiences (Psychoticism). All items are on a four-point scale in which higher scores indicate more severe psychopathology. A sample of 818 adults completed the PID-5, with sampling weighted to be representative of the U.S. population in terms of gender, age, race/ethnicity, and education. Details of sampling methods and demographics are described in Krueger et al. (2012).

In addition to rating themselves, respondents rated two vignettes per trait, for a total of ten vignettes. The two vignettes were written so as to reflect average and elevated ranges of the trait. Posterior estimates of vignette severities, v, confirmed that the vignettes were rated in the intended range. Respondents rated each vignette using three items from the

informant report version of the PID-5 (Markon, Quilty, Bagby, & Krueger, 2013). Informant report items were written to be as similar to the self-report items as possible. The complete set of vignettes and the items used to rate them can be found in Appendix A.

Lastly, respondents completed an eight item measure from the The Patient-Reported Outcomes Measurement Information System (PROMIS) assessing their ability to participate in social roles (PROMIS Ability to Participate in Social Roles and Activities – Short Form 8a). Reliability for this measure is high (>.98), and item-total correlations are acceptable (.65-.85) in a sample drawn from the general population (Hahn et al., 2010). Items follow a five-point Likert response format. Due to software constraints at the time of these analyses, only the 553 individuals with complete response data for the five traits, ten vignettes, and PROMIS were included.

Data Simulation.

All data was simulated in R (Team, 2010, version 3.1.2). Simulated data, like the observed data, consisted of 553 respondents with five traits each, rating themselves with 37 items with a four-point response format. To investigate the effect of the number of vignettes and vignette ratings on response style parameter estimation, data sets were simulated from the b_i model fit to the observed data. The number of simulated vignettes varied (v = 1, 3, 5 or 10) as did the number of items used to rate them (j = 1, 2, or 3). Each condition was replicated ten times.

For the remaining simulations, parameters were drawn from the observed correlation matrix of the response style parameters and content trait parameters (reported in Appendix B). To test the effect of correlations between response style and content traits, five variants of the β_i model were simulated, in which the correlation between $\boldsymbol{\theta_i}$ and b_i was set to either the observed correlations, 0.0, 0.25, 0.5, or 0.75. Ten simulated data sets were drawn for each condition, and the GRM and data-generating model were fit to each data set.

Priors & Estimation Procedures.

All models were estimated using RStan (*RStan: the R interface to Stan, Version 2.8.0*, 2015). Appendix C includes Stan code for the GRM and $a_i + b_i$ models. Appendix D describes priors and convergence diagnostics.

Metrics of Learning and Model Fit.

Bayesian learning was assessed by calculating Kullback-Liebler (KL) divergence between prior and posterior distributions, as defined in Equation 1. Since the closed-form of the posterior distribution is unknown, $p(\theta|x)$ was defined by cubic smoothing splines fit to the empirical probability density of the posterior draws. Model fit was assessed via WAIC (Widely Applicable Information Criterion; Watanabe, 2010, Dec). WAIC approximates leave-one-out cross validation, which in turn approximates how well the model will predict new data. WAIC is based on the same deviance scale as the Deviance Information Criterion and Bayesian Information Criterion, in which lower values indicate better model fit. Models were also compared in terms of their ability to predict observed data–i.e., via posterior predictive checks (A. Gelman, Goegebeur, Tuerlinckx, & Van Mechelen, 2000). This was operationalized as the probability that the response data predicted by model parameter estimates at a given draw equaled the observed data, averaged over draws. A PPC_y value of 1 would indicate a model that correctly predicted the observed data 100% of the time.

Results

Information Gained from Vignettes

Figure 4 depicts KL divergence of the b_i parameter as a function of the number of vignettes rated, and the number of items rated per vignette. Values of 0 mean no information has been gained. Even with one vignette, rated once, KL divergence was greater than 0. As the number of vignettes and ratings increased, KL divergence increased. Generally, it appears as though five to six vignette ratings per trait increased the amount of information learned about b_i beyond baseline conditions. As an auxiliary analysis, we

also assessed how modeling response styles using vignettes affected KL divergence for other parameters in PS IRT models (Table E1; Appendix E). All models increased the amount of information learned about the traits of interest, θ_i.

Model Accuracy & Comparison in Simulated Data

***b_i* data.** Fit statistics for the b_i and GRM models in data simulated to reflect ARS/DRS are reported in rows 1-5 of Table 1. Regardless of the correlation between content traits $\boldsymbol{\theta_i}$ and response style b_i, WAIC consistently favored the correct, data-generating model. The posterior predictive check was equivocal. Across all conditions and parameters, the b_i model resulted in lower RMSE for both person and item parameters. In all but one condition, the b_i model also resulted in lower parameter bias. In sum, when yea-saying and nay-saying response styles are present, the b_i outperforms the GRM in nearly all measures of model fit and accuracy.

***a_i* data.** Fit statistics of the a_i model and GRM in data simulated to reflect inattention or lack of traitedness are reported in row 6 of Table 1. The PS IRT model was preferred by all measures of model fit and accuracy, halving the bias and RMSE of test parameters.

***$a_i + b_i$* data.** Fit statistics of the $a_i + b_i$ model and GRM in data simulated to reflect attentiveness/traitedness and ARS/DRS are reported in row 7 of Table 1. Again, the PS IRT model was preferred by nearly all fit indices, demonstrating better parsimony-adjusted fit, more accurate prediction of observed response data, and lower RMSE values across almost all parameters.

***ω_i* data.** Model fit statistics for the ω_i model and GRM in data simulated to reflect ERS and MRS are reported in row 8 of Table 1. In this case, WAIC favored the GRM. The PS IRT model, therefore, was not necessarily favored in data with ERS and MRS.

***$\omega_i + b_i$* data.** Last, model fit statistics for the $\omega_i + b_i$ model and GRM in data simulated to reflect both ERS/MRS and ARS/DRS are reported in the last row of Table 1. Again, fit statistics did not conclusively favor one model over the other. The PS IRT model

resulted in more accurate and unbiased parameter estimates, and was slightly more effective in predicting the observed data. It was not, however, parsimonious.

Model Performance in Observed Personality Data and Criterion Validity

Table reftab:pid reports fit of the PS IRT models and GRM to observed personality data. All PS IRT models outperformed the GRM. The best-fitting model was the $a_i + b_i$ model, indicating responses may have been particularly influenced by inattention or traitedness, and ARS/DRS. Models were roughly equivalent in predicting observed response data, with the $\omega_i + b_i$ model slightly outperforming the others.

Figure 5 depicts the effect of estimating ω_i model for one respondent. This individual had a midpoint response style, rating all 10 vignettes as a 2 or 3 on the 4-point scale. Their ω_i estimate was 3.01, meaning they perceived category thresholds as three times as far apart than the average respondent. As a result of this pronounced response style, their endorsement of the highest response option in rating their own degree of Manipulativeness shifted the posterior estimate of that trait significantly upward. Figure S1 shows how estimating response style changes shifts personality estimates in the sample at large.

We also compared the five PS IRT models in terms of their ability to predict self-reported social function. In general, the incremental improvement in prediction was small ($\Delta R^2 = 0.01$; total $R^2 = 0.28$), but in the case of the ω_i model, the change in R^2 was statistically significant ($\Delta R^2 = 0.02$; total $R^2 = 0.29$). In sum, correcting for response style reveals trait variance that predicts social function.

Discussion

This study had three aims: One, to assess the number of vignette rating needed to identify response style parameters; Two, to integrate and compare multiple ways of modeling vignette data that have been proposed in the literature, but never directly compared, and to test the effect of response style and content trait correlations on model performance; Three, to implement response style models in self-reported personality data, and assess their ability to fit the response data and predict social dysfunction.

These simulations revealed response style parameters can be estimated with as little as one vignette, rated once. Increasing either the number of vignettes or the number of ratings per vignette increases the amount of information gained–although it appears more efficient to increase the number of ratings per vignette than the number of vignettes. Furthermore, modeling response style increases the amount of information gained about other parameters in the model. PS IRT models applied to vignette ratings are therefore useful to both test developers, as a way to maximize information about item parameters, and test administrators, as a way of learning more about person parameters.

The second aim of the study revealed that vignette-based PS IRT models are effective in capturing response styles, when present in the data. In particular, models of yea-saying and nay-saying (the b_i model), inattention (a_i), and combinations of the two were both more accurate and more parsimonious than the GRM. In addition, the performance of the b_i model was unaffected by the degree to which response style and content traits were correlated. This illustrates the importance of using vignettes to estimate response style. It is possible to imagine a case where response style is perfectly correlated with content traits (the degree to which ARS is a linear function of Neuroticism, for example). Without vignettes, it would be impossible to identify the unique contributions of Neuroticism and ARS to self-report data. Vignette ratings, however, reflect only ARS and the vignette severity, and in doing so allow one to parse the variance attributable to the two components.

Unexpectedly, models of ERA and MRS (the ω_i model) were generally preferred by measures of absolute fit, but not by parsimony-adjusted fit indices. Modeling ERS and MRS improved the absolute error in predicting response data, and improved the accuracy of some item parameters, but this improvement was outweighed by the increased complexity of the response style model. Since the degree of simulated ERS and MRS was based on the observed personality data, it may be that the models would perform better in contexts more strongly influenced by ERS and MRS. This would be inconsistent with

Wetzel, Carstensen, and Böhnke (2013), which shows ERS and MRS to be a significant factor in personality data. However, other analyses have shown that ERS and MRS are less significant predictor of item functioning than other factors, such as gender (Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013). In addition, the PID-5's four-point response format may have dampened MRS, explaining this pattern of results.

Finally, vignette ratings revealed response styles were sufficiently pronounced in responses to the Personality Inventory for DSM-5 that all of the proposed models were superior to the GRM. This suggests that response styles are a very important component of self-report personality data. Furthermore, the fact that $a_i + b_i$ model fit best suggests that inattentiveness (or lack of traitedness) and yea-saying/nay-saying were the most prevalent response styles in this context. And yet, the PS IRT models improved criterion validity by only a small degree. In one case (the ω_i model), the increase was statistically significant, but across models the magnitude of the effect was of questionable practical utility. There are a number of possible explanations for this finding. For one, in the hypothetical case in which yea-saying and neuroticism are perfectly correlated, the estimate of neuroticism from the GRM would equal the estimate of neuroticism from the PS IRT model, plus b_i. The fact that neuroticism and b_i can be distinguished will not change the fact that the variance in social function accounted for by each will be equally well accounted for by the GRM estimate. In the observed data, correlations between b_i and personality traits were moderate to large, meaning that PS IRT parameters may not predict much variance in social function that is not encapsulated in GRM estimates. It may be that criterion validity will only increase when b_i and content traits are less closely correlated, although this contradicts the observations of Plieninger (2017), who found the effect of response style was most prominent when correlations between content traits and response styles were high. A second explanation, at least in the case of models with parameters that change the scale of the category response function (a_i and ω_i), is that response styles add noise to the relationship between trait and outcome, and removing the

source of noise cannot recover the lost information about the trait. An intriguing potential use for these parameters, however, is as weights. Specifically, a_i might conceivably be used as a way to down-weight data from inattentive respondents.

A third possible explanation for the lack of incremental validity—and this study's primary limitation—is that the criterion variable is also affected by response style. We chose not to correct the criterion data for response style in order to created a more stringent test of prediction. It is more compelling that response style-corrected estimates of personality pathology predict *uncorrected* estimates of social function, given that the same response styles are likely influencing both. However, whether response style parameters can be shown to moderate the relationship between personality and criterion variables that are both corrected for response style, or criterion variables that are not dependent on self-report, is an empirical question that can be addressed in future research.

Conclusions

Adding as little as one vignette rating to a self-report measure allows for the estimation of response style parameters. Modeling response styles using vignettes generally improves model fit, and increases the amount of information learned about both person and test parameters. In self-reported personality data, all response styles were prevalent enough that explicitly estimating them improved model fit, although the best-fitting model incorporated just inattention and yea-saying/nay-saying. Furthermore, using vignettes improves the prediction of social dysfunction, but only to a small degree. In sum, the current research supports the use of response style models not only by both test developers and test administrators.

References

Baumgartner, H. & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, *38*(2), 143–156.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*(4), 381–409.

Bolt, D. M. & Johnson, T. R. [Timothy R]. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352.

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

Dolnicar, S. & Grün, B. (2009, August 1). Response style contamination of student evaluation data. *Journal of Marketing Education*, *31*(2), 160–172. doi:10.1177/0273475309335267

Falk, C. F. & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. doi:10.1037/met0000059

Ferrando, P. J. (2014, July 4). A factor-analytic model for assessing individual differences in response scale usage. *Multivariate Behavioral Research*, *49*(4), 390–405. doi:10.1080/00273171.2014.911074

Gelman, A. [A.], Goegebeur, Y., Tuerlinckx, F., & Van Mechelen, I. (2000, January 1). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(2), 247–268. doi:10.1111/1467-9876.00190

Gelman, A. [Andrew] & Rubin, D. B. (1992, November). Inference from iterative simulation using multiple sequences. *7*(4), 457–472. doi:10.1214/ss/1177011136

Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011, June 1). Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, *52*(2), 246–261. doi:10.1177/0022146510396713

Hahn, E. A., DeVellis, R. F., Bode, R. K., Garcia, S. F., Castel, L. D., Eisen, S. V., ... Cella, D., et al. (2010). Measuring social health in the patient-reported outcomes measurement information system (PROMIS): Item bank development and testing. *Quality of Life Research*, *19*(7), 1035–1044.

He, J., Bartram, D., Inceoglu, I., & Van de Vijver, F. J. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology*, *45*(7), 1028–1045.

Herk, H. v., Poortinga, Y. H., & Verhallen, T. M. M. (2004, May 1). Response styles in rating scales evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. doi:10.1177/0022022104264126

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of applied psychology*, *75*(5), 581.

Jin, K.-Y. & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138.

Johnson, T. R. [Timothy R.]. (2003, December 1). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *68*(4), 563–583. doi:10.1007/BF02295612

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, *98*(1), 191–207.

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2012). Initial construction of a maladaptive personality trait model and inventory for DSM-5. *Psychological medicine*, *42*(9), 1879–1890.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, *22*(1), 79–86.

Markon, K. E., Quilty, L. C., Bagby, R. M., & Krueger, R. F. (2013). The development and psychometric properties of an informant-report form of the personality inventory for DSM-5 (PID-5). *Assessment*, *20*(3), 370–383.

McCrae, R. R., Costa, P. T., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B. (1989, February). A caution on the use of the MMPI k-correction in research on psychosomatic medicine. *51*(1), 58–65.

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. doi:10.1037/a0019216

Moors, G. (2012, April 1). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298. doi:10.1080/1359432X.2010.550680

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2011, August 1). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *41*(1), 13–47. doi:10.1111/j.1467-9531.2011.01238.x

Mõttus, R., Allik, J., Realo, A., Rossier, J., Zecca, G., Ah-Kion, J., ... Johnson, W. (2012, June 27). The effect of response style on self-reported conscientiousness across 20 countries. doi:10.1177/0146167212451275

Plieninger, H. (2017, January 1). Mountain or molehill? a simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*(1), 32–53. doi:10.1177/0013164416636655

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001, March 1). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, *96*(453), 20–31. doi:10.1198/016214501750332668

RStan: The r interface to stan, version 2.8.0. (2015). Retrieved from http://mc-stan.org/rstan.html

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4), 100.

Team, R. D. C. (2010). R: A language and environment for statistical computing (Version 2.11.1). Vienna, Austria. Retrieved from http://www.R-project.org

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*(1), 96.

Wetzel, E., Böhnke, J. R., Carstensen, C. H., Ziegler, M., & Ostendorf, F. (2013). Do individual response styles matter? assessing differential item functioning for men and women in the NEO-PI-r. *Journal of Individual Differences*, *34*(2), 69.

Wetzel, E., Böhnke, J. R., & Rose, N. (2016, April 1). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement*, *76*(2), 304–324. doi:10.1177/0013164415591848

Wetzel, E. & Carstensen, C. H. [Claus H]. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*.

Wetzel, E., Carstensen, C. H. [Claus H.], & Böhnke, J. R. (2013, April 1). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178–189. doi:10.1016/j.jrp.2012.10.010

Xie, Y. & Carlin, B. P. (2006, October 1). Measures of bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, *136*(10), 3458–3477. doi:10.1016/j.jspi.2005.04.003

Table 1

Accuracy of PS IRT and GRM models as a function of the correlation between response style and content traits

			RMSE			Bias		
	WAIC	PPC_y	θ_i	a_j	b_{jk}	θ_i	a_j	b_{jk}
$corr(b_i, \theta_i) = 0$								
GRM	64,124.80	0.56	0.53	0.23	0.67	-0.21	0.07	-0.63
b_i	**61,293.82**	**0.57**	**0.40**	**0.20**	**0.66**	**-0.19**	**-0.05**	**-0.62**
$corr(b_i, \theta_i) = 0.25$								
GRM	61,472.77	**0.57**	0.50	0.33	0.71	-0.21	0.24	**-0.65**
b_i	**58,927.87**	**0.57**	**0.40**	**0.24**	**0.70**	**-0.19**	**-0.08**	**-0.65**
$corr(b_i, \theta_i) = 0.50$								
GRM	60,851.20	**0.59**	0.47	0.52	0.74	-0.20	0.46	-0.69
b_i	**57,953.93**	**0.59**	**0.40**	**0.25**	**0.72**	**-0.18**	**-0.07**	**-0.69**
$corr(b_i, \theta_i) = 0.75$								
GRM	61,151.4	**0.59**	0.42	0.61	**0.73**	-0.20	0.57	**-0.69**
b_i	**58,248.21**	**0.59**	**0.39**	**0.23**	**0.73**	**-0.19**	**-0.11**	-0.70
$corr(b_i, \theta_i) =$ observed								
GRM	61,550.94	0.58	0.48	0.44	0.73	-0.21	0.38	-0.68
b_i	**58,772.10**	**0.59**	**0.41**	**0.22**	**0.72**	**-0.19**	**-0.10**	**-0.68**
a_i								
GRM	71,818.47	0.53	0.67	0.50	0.55	-0.27	0.40	-0.40
a_i	**61,909.85**	**0.54**	**0.50**	**0.28**	**0.26**	**-0.09**	**-0.23**	**-0.18**
$a_i + b_i$								
GRM	73,653.91	0.51	0.74	0.47	0.37	-0.19	0.21	**-0.05**
$a_i + b_i$	**63,483.22**	**0.53**	**0.44**	**0.36**	**0.23**	**-0.04**	**-0.19**	-0.16
ω_i								
GRM	**62,376.64**	0.58	**0.44**	**0.34**	0.54	**-0.04**	**-0.02**	-0.43
ω_i	64,695.88	**0.59**	**0.44**	0.36	**0.21**	-0.07	-0.06	**-0.06**
$\omega_i + b_i$								
GRM	**62,024.15**	0.58	0.64	0.82	0.66	0.10	-0.73	-0.60
$\omega_i + b_i$	64,048.39	**0.60**	**0.42**	**0.32**	**0.22**	**-0.08**	**-0.13**	**-0.13**

Note. Fit of GRM and PS IRT models in data simulated to reflect ARS and DRS, with the correlation between response style and content traits varying, (rows 1-5), inattention (row 6), inattention plus ARS/DRS (row 7), ERS/MRS (row 8), and ERS/MRS plut ARS/DRS (row 9). PPC = Posterior Predictive Check; RMSE = Root Mean Square Error; WAIC = Widely Applicable Information Criterion.

Table 2

Fit of GRM and PS IRT models to observed data

	WAIC	PPC_y
GRM	64,349.65	0.59
b_i	60,915.17	0.60
a_i	60,050.30	0.59
$a_i + b_i$	**57,496.51**	0.60
ω_i	61,559.62	0.60
$\omega_i + b_i$	59,401.71	**0.61**

Note. Fit of the GRM and PS IRT models to self-reported personality data from the PID-5.

PPC_y = Posterior Predictive Check; WAIC = Widely Applicable Information Criterion.

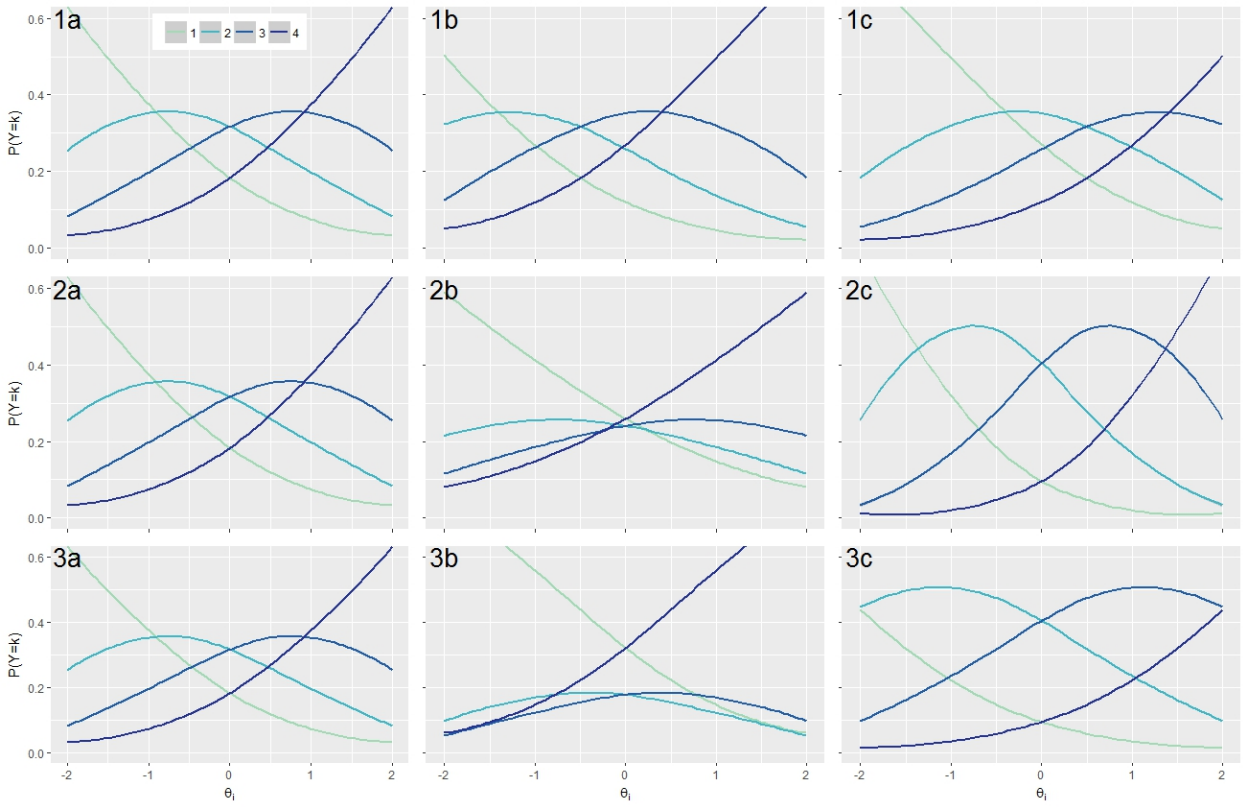

Figure 1-3. Figure 1 depicts category response curves for an item in the b_i model when (1a) $b_i = 0$, (1b) $b_i < 0$, an acquiescent response style, and (1c) $b_i > 0$, a disacquiescent response style. Figure 2 depicts the a_i model when (2a) $a_i = 1$, (2b) $a_i < 1$, reflecting inattentiveness, and (2c) $a_i > 1$, reflecting high attentiveness or traitedness. Figure 3 depicts category response curves for an item in the ω_i model when (3a) $\omega_i = 1$, (3b) $\omega_i < 1$, reflecting extreme response style, and (3c) $\omega_i > 1$, reflecting a midpoint response style.

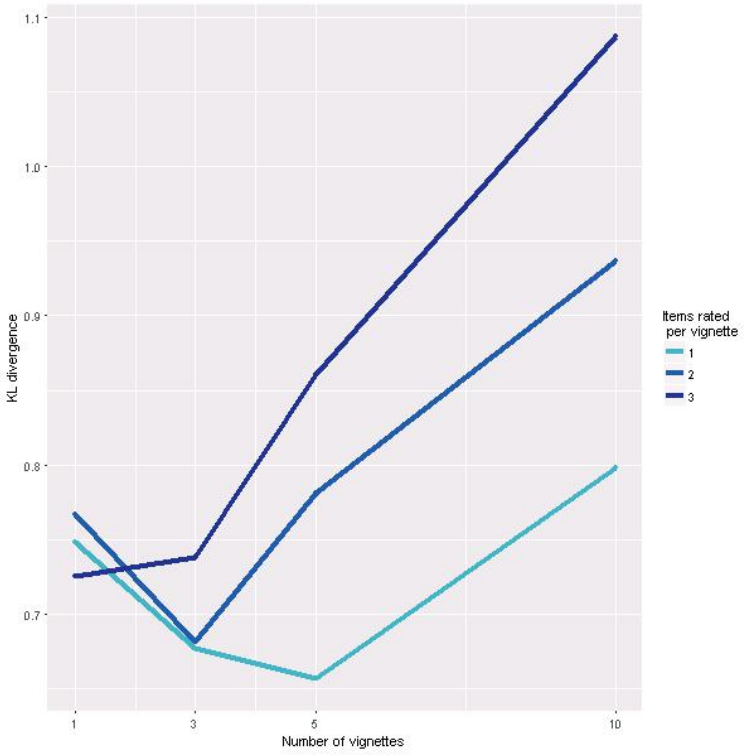

Figure 4. KL divergence of b_i as a function of the number of vignettes, and number of items rated per vignette.

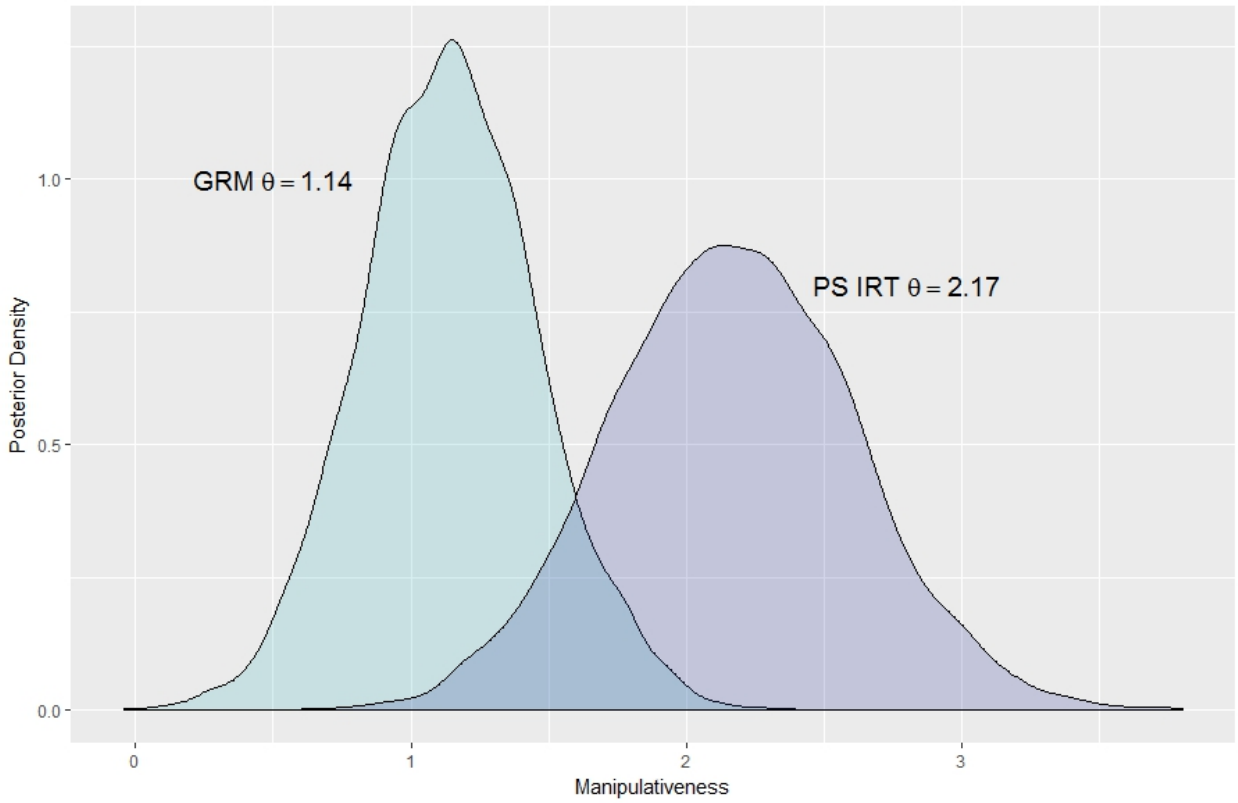

Figure 5. Posterior density of estimates from the graded response model (GRM) and the ω_i model for an individual with a midpoint response style ($\omega_i = 3.01$). Using a 4-point scale, this respondent rated all vignettes as either 2 or 3, but in responding to the item "I can certainly turn on the charm if I need to get my way," rated his or herself as a 4.

Appendix A

Vignettes and Items Used for Rating[3]

Emotional Lability

Erica gets upset very easily. Small setbacks, like finding a button missing on her coat, can frustrate her so much that she is unable to leave the house on time. Her friends are hesitant to go with her to public events, because they never know how she will react to new people and situations, and she sometimes blows up at people she has just met. Erica's changeable emotions have prevented her from maintaining close relationships with others.

When she was working, Maria was almost always calm, and became upset by customers only a few times. When business was slow and her position was cut, Maria became worried and moody. Despite being sad about losing her job, she revised her resume and was looking for a new job the next day. Some things can unexpectedly change Maria's mood for the worse, but her friends feel they can count on her to return to her usual self pretty quickly.

62. . . . has much stronger emotional reactions than almost everyone else.

102. . . . is a highly emotional person.

181. . . . has unpredictable emotions.

[3]Item numbers correspond to the PID-5 Informant-Report Form.

Withdrawal

Lucas prefers not to interact with other people and has few relationships. When he goes to dinner in his college's dining hall he sits alone, and if other students sit down at the same table he avoids talking to them. Lucas initially wanted to go into banking, but his classes required him to work with others in group projects, so he switched to math. He hardly ever has to work with other students in math classes.

Peter calls his friends and girlfriends frequently, but sometimes doesn't feel like talking to others. He prefers to be with others, but is fine going on errands and trips shopping by himself. If his roommates aren't around after work, Peter sometimes calls people he knows to see what they're doing. He likes parties and social gatherings, and sometimes has get-togethers with groups of friends. Sometimes, though, he just wants to be by himself to relax.

82. . . . keeps their distance from people.

136. . . . doesn't like spending time with others.

182. . . . doesn't deal with people unless they have to.

Manipulativeness

Seth is very good at reading other people. Within a few minutes of being introduced, he can figure out what makes someone "tick." He enjoys learning about other people, so that he can use their weaknesses to his own benefit. By flattering and praising his boss, Seth has manipulated office politics to his own advantage, resulting in the demotion of a colleague he dislikes and his own promotion.

Natalie keeps up some relationships, even though she doesn't feel close to the other person, mainly because it's good for her work. She knows her good friends and boyfriends care about her and feel close to her, and she feels the same way. Sometimes, though, she wonders if she's taking advantage of them. She doesn't hesitate to end a relationship if it's not in her best interests in the long run. Although sometimes she has misgivings about falling out of touch with people who aren't useful to her, in general she doesn't have any regrets about doing so.

107. . . . is good at making people do what they want them to do.

180. . . . can certainly turn on the charm if they need to get their way.

219. ...finds it is easy to take advantage of others.

Irresponsibility

Although Ellen promised she would attend the wedding of a childhood friend, on the weekend of the wedding she just didn't feel like going, so she drove to the city instead and spent the weekend in clubs and theaters. Later, she realized that her roommate's shoes, which Ellen had borrowed for the wedding, were stained and scuffed. She couldn't replace them, since she had spent all her money over the weekend. She just slipped the shoes back into her roommate's closet as they were, hoping the damage wouldn't be noticed.

Harold has been employed at the same store for multiple years. The manager likes Harold, even though Harold has been late to work sometimes, and occasionally puts away store products incorrectly. Harold often works late to make sure a job gets done, but sometimes after working late he's tired and "forgets" to pick up things he promised his wife he'd get, like groceries. Harold recently went on a vacation with his wife to take a break from work, but only told the manager about the sick relative they visited.

129. . . . is often pretty careless with their own and others' things.

156. . . . seems to make promises that they don't intend to keep.

201. . . . just skips appointments or meetings if they are not in the mood.

Unusual Beliefs and Experiences

Even though he and his friends are the only people in their workplace breakroom, Neil says that there are other people sitting at the next table, and they're saying bad things about his department. Neil is sure his co-workers are using magnets to make his cash register add up numbers incorrectly, so that he'll be fired. He is worried that the residual magnetism from the cash register may transfer itself to him, so he wears gloves and a heavy coat while at work.

Sonia is a scientist, and skeptical about most things that haven't been proven through experimentation. She likes science fiction books and movies about ghosts, though, and believes that people really never know what happens in the afterlife. Sonia is confident that everything has a reason, and believes that, in time, logical explanations are found for most mysterious phenomena. However, she has had a few experiences that are difficult to explain. For example, she once thought she might have heard the voice of her grandmother who passed away recently.

99. . . . sometimes hears things that aren't really there.

139. . . . has seen things that weren't really there.

194. . . . often makes unusual connections between things.

Appendix B

Table B1

Correlations of PS IRT response parameter and trait estimates

	EL	Irr	Man	UBE	With
EL	1.00	-	-	-	-
Irr	0.77	1.00	-	-	-
Man	0.58	0.73	1.00	-	-
UBE	0.70	0.80	0.74	1.00	-
With	0.65	0.72	0.48	0.59	1.00
b_i	0.41	0.62	0.49	0.59	0.46
a_i	0.03	-0.11	-0.01	-0.11	-0.03
$a_i + b_i$					
a_i	0.11	-0.01	0.04	-0.03	-0.10
b_i	0.43	0.56	0.51	0.58	0.47
ω_i	0.06	0.10	0.32	0.12	0.05
$\omega_i + b_i$					
ω_i	-0.10	-0.11	0.15	-0.05	-0.06
b_i	0.41	0.64	0.44	0.59	0.46

Note. Observed Pearson correlations between content and response style traits, which were used to generate simulated data. EL=Emotional Lability, Irr=Irresponsibility, Man=Manipulativeness, UBE=Unusual Beliefs and Experiences, With=Withdrawal.

Appendix C

Model Code

GRM

```
library(rstan)

## format of data

# y is a N by M matrix of item responses
# N = number of respondents
# M = number of responses
# V = number of vignettes
# J = number of items (M != J because some items are used in both self
   and vignette ratings)
# K = number of categories per item
# F = number of facets (latent traits)
# jm = an indexing vector linking item j to response m
# tim = an indexing vector linking trait i (the trait estimate for
   individual i) to response m
# tvm = an indexing vector linking trait v (the trait estimate for
   vignette v) to response m

data <- list(y = y, M = M, N = N, V = V, J = J, K = K, F = F, jm = jm,
   tim = tim, tvm=tvm)

## specify model

model_code <- '
  data {
```

```
22    int<lower=1> M; // number of responses
23    int<lower=1> N; // number of respondents
24    int<lower=1> y[N,M]; // response data
25    int<lower=1> V; // number of vignettes
26    int<lower=1> J; // number of items
27    int<lower=2> K; // number of response options
28    int<lower=2> F; // number of facets
29    int<lower=1,upper=J> jm[M]; // vector indexing item j to response m
30    int<lower=1,upper=F> tnm[J]; // vector indexing response m to
         theta_i
31    int<lower=1,upper=V> tvm[M-J]; // vector indexing response m to
         theta_v
32    }
33  parameters {
34    matrix<lower=-5,upper=5>[N,F] theta_n_raw;
35    vector<lower=-5,upper=5>[V] theta_v_raw;
36    corr_matrix[F] theta_sigma;
37    real<lower=0,upper=5> alpha_j[J];
38    ordered[K-1] beta_jk_raw[J-1];
39    real<upper=0> beta_jk_1_raw;
40    real<lower=0> beta_jk_3_raw;
41    }
42  transformed parameters {
43    ordered[K-1] beta_jk[J];
44    vector[F] theta_mu;
45    vector<lower=0>[F] theta_sd;
46    matrix[F,F] theta_cov_decomp; // cholesky decomposition of the
         covariance matrix between factors
47    for (d in 1:F) {
```

```
48      theta_mu[d] = 0;
49      theta_sd[d] = 1;
50    }
51    theta_cov_decomp = diag_matrix(theta_sd) * cholesky_decompose(
          theta_sigma);
52    beta_jk[1,1] = beta_jk_1_raw; beta_jk[1,2] = 0; beta_jk[1,3]<-
          beta_jk_3_raw;
53    for (j in 2:J) {
54      beta_jk[j] = beta_jk_raw[j-1];
55      }
56    }
57  model {
58    matrix[N,M] theta;
59    alpha_j ~ lognormal(0,1);
60    beta_jk_1_raw ~ normal(0,1);
61    beta_jk_3_raw ~ normal(0,1);
62    for (j in 1:(J-1)) {
63      beta_jk_raw[j] ~ normal(0,1);
64      }
65    theta_v_raw ~ normal(0,1);
66    for (n in 1:N) {
67      theta_n_raw[n] ~ multi_normal_cholesky(theta_mu, theta_cov_decomp
            );
68      for (m in 1:J) {
69        theta[n,m] = theta_n_raw[n,tnm[m]];
70        }
71      for (m in J+1:M) {
72        theta[n,m] = theta_v_raw[tvm[m-J]];
73        }
```

```
      for (m in 1:M) {
        y[n,m] ~ ordered_logistic(alpha_j[jm[m]]*theta[n,m], beta_jk[jm
            [m]]);
      }
    }
  }
'

# specify initial values

initvals = list(theta=rep(0,N+V), alpha_j=rep(1,J), beta_jk_raw=matrix(
   nrow=(J-1), ncol=(K-1), c(-1,0,1), byrow=T),
beta_jk_1_raw=-1, beta_jk_3_raw=1, theta_n_raw=matrix(nrow=N, ncol=F,
   0), theta_v_raw=rep(0,V))
initfun <- function(x) { return(initvals); }

# compile model

model_fit1 <- stan(model_code = model_code, data = data, iter = 1,
   chains = 1, init = initfun)

# run three chains in parallel

chains <- 3
cl <- makeCluster(chains, type = "SOCK")
clusterExport(cl, list=c("model_fit1", "chains", "data", "initfun", "
   initvals"))
clusterEvalQ(cl, library(rstan))
model_fit <- parLapply(cl, 1:chains, fun = function(chain) {
```

```
                stan(fit = model_fit1, seed=123, data = data, chains =
                   1, warmup=500, iter=3000, verbose = TRUE, init =
                   initfun, refresh = -1, chain_id=chain)
            })
stopCluster(cl)

model_fit <- sflist2stanfit(model_fit)
```

$a_i + b_i$ Model Code

```
library(rstan)

## format of data
# y is a N by M matrix of item responses
# N = number of respondents
# M = number of responses
# V = number of vignettes
# J = number of items (M != J because some items are used in both self and vignette ratings)
# K = number of categories per item
# F = number of facets (latent traits)
# jm = an indexing vector linking item j to response m
# tim = an indexing vector linking trait i (the trait estimate for individual i) to response m
# tvm = an indexing vector linking trait v (the trait estimate for vignette v) to response m

data <- list(y = y, M = M, N = N, V = V, J = J, K = K, F = F, jm = jm, tim = tim, tvm=tvm)

## specify model

model_code <- '
  data {
    int<lower=1> M; // number of responses
    int<lower=1> N; // number of respondents
    int<lower=1> y[N,M]; // response data
```

```
24    int<lower=1> V; // number of vignettes
25    int<lower=1> J; // number of items
26    int<lower=2> K; // number of response options
27    int<lower=2> F; // number of facets
28    int<lower=1,upper=J> jm[M]; // vector indexing item j to response m
29    int<lower=1,upper=F> tnm[J]; // vector indexing response m to
         theta_i
30    int<lower=1,upper=V> tvm[M-J]; // vector indexing response m to
         theta_v
31    }
32  parameters {
33    matrix<lower=-5,upper=5>[N,F+2] theta_n_raw;
34    vector<lower=-5,upper=5>[V] theta_v_raw;
35    corr_matrix[F+2] theta_sigma;
36    real<lower=0,upper=5> alpha_j[J];
37    ordered[K-1] beta_jk_raw[J-1];
38    real<upper=0> beta_jk_1_raw;
39    real<lower=0> beta_jk_3_raw;
40    }
41  transformed parameters {
42    ordered[K-1] beta_jk[J];
43    vector[F+2] theta_mu;
44    vector<lower=0>[F+2] theta_sd;
45    matrix[F+2,F+2] theta_cov_decomp; // cholesky decomposition of the
         covariance matrix between factors
46    for (d in 1:F+2) {
47      theta_mu[d] = 0;
48      theta_sd[d] = 1;
49    }
```

```
50    theta_cov_decomp = diag_matrix(theta_sd) * cholesky_decompose(
         theta_sigma);
51    beta_jk[1,1] = beta_jk_1_raw; beta_jk[1,2] = 0; beta_jk[1,3] =
         beta_jk_3_raw;
52    for (j in 2:J) {
53      beta_jk[j] = beta_jk_raw[j-1];
54      }
55    }
56  model {
57    matrix[N,M] theta;
58    real alpha_i[N];
59    real beta_i[N];
60    alpha_j ~ lognormal(0,1);
61    beta_jk_1_raw ~ normal(0,1);
62    beta_jk_3_raw ~ normal(0,1);
63    for (j in 1:(J-1)) {
64      beta_jk_raw[j] ~ normal(0,1);
65      }
66    theta_v_raw ~ normal(0,1);
67    for (n in 1:N) {
68      theta_n_raw[n] ~ multi_normal_cholesky(theta_mu, theta_cov_decomp
           );
69      alpha_i[n] = exp(theta_n_raw[n,6]);
70      beta_i[n] = theta_n_raw[n,7];
71      for (m in 1:J) {
72        theta[n,m] = theta_n_raw[n,tnm[m]];
73        }
74      for (m in J+1:M) {
75        theta[n,m] = theta_v_raw[tvm[m-J]];
```

```
        }
      for (m in 1:M) {
        y[n,m] ~ ordered_logistic(alpha_i[n]*alpha_j[jm[m]]*theta[n,m],
            beta_jk[jm[m]]+beta_i[n]);
      }
    }
  }
'

# specify initial values

initvals = list(theta=rep(0,N+V), alpha_j=rep(1,J), beta_jk_raw=matrix(
   nrow=(J-1), ncol=(K-1), c(-1,0,1), byrow=T),
beta_jk_1_raw=-1, beta_jk_3_raw=1, theta_n_raw=matrix(nrow=N, ncol=F,
   0), theta_v_raw=rep(0,V), alpha_i=rep(1,N), beta_i=rep(0,N))
initfun <- function(x) { return(initvals); }

# compile model

model_fit1 <- stan(model_code = model_code, data = data, iter = 1,
   chains = 1, init = initfun)

# run three chains in parallel

chains <- 3
cl <- makeCluster(chains, type = "SOCK")
clusterExport(cl, list=c("model_fit1", "chains", "data", "initfun", "
   initvals"))
clusterEvalQ(cl, library(rstan))
```

```
model_fit <- parLapply(cl, 1:chains, fun = function(chain) {
                stan(fit = model_fit1, seed=123, data = data, chains =
                   1, warmup=500, iter=3000, verbose = TRUE, init =
                   initfun, refresh = -1, chain_id=chain)
             })
stopCluster(cl)

model_fit <- sflist2stanfit(model_fit)
```

Appendix D

Priors and Convergence Diagnostics

Item discrimination parameters were given $\text{Log}\mathcal{N}(0, 1)$ priors. Item-threshold parameters were given $\mathcal{N}(0, 1)$ priors (excepting the second threshold of the first item, which was set to 0 to fix the location of the model). Vignette trait values were given $\mathcal{N}(0, 1)$ priors. Person parameters were given a multivariate normal prior with with $\mathcal{N}(\boldsymbol{\mu} = 0, \boldsymbol{\Sigma})$, with the diagonal elements (variances) of $\boldsymbol{\Sigma}$ set to 1, and off-diagonal elements (covariances) given a $\text{U}(-1, 1)$ prior.

Three chains were run for each model, each beginning with 500 warm-up iterations that were discarded, followed by 2500 iterations that were retained for analysis. Model convergence was assessed by visual examination of trace plots and posterior distributions for all item parameters and a random subset of person parameters. Convergence was also monitored using Gelman and Rubin's potential scale reduction factor $\hat{R}$ (Andrew Gelman & Rubin, 1992). Values nearer to 1 are preferred, and although there is no strict cutoff for $\hat{R}$, values below 1.1 can be considered satisfactory. In all models, $\hat{R}$ was less than 1.05 after 3000 iterations.

Appendix E

KL Divergence

Table E1

Kullback-Leibler divergence between prior and posterior parameter distributions

	a_j	b_{jk}	v	θ_i	b_i	a_i	ω_i
GRM	2.29	3.85	2.68	0.68		-	-
b_i	2.62	4.13	2.66	1.09	1.11	-	-
a_i	2.01	3.77	3.17	1.07	-	1.58	-
$a_i + b_i$	2.25	3.04	1.99	1.22	1.08	1.36	-
ω_i	2.45	1.69	2.61	1.12	-	-	1.70
$\omega_i + b_i$	3.19	1.39	1.92	1.24	1.32	-	1.59

Note. Values are means of sets of parameters. Kullback-Leibler divergences of zero mean the prior and posterior distributions are equal, and no information has been learned from the data. Values greater than zero reflect the degree to which the data has changed the posterior relative to the prior, with larger values reflecting a greater change.

Appendix F

Bias

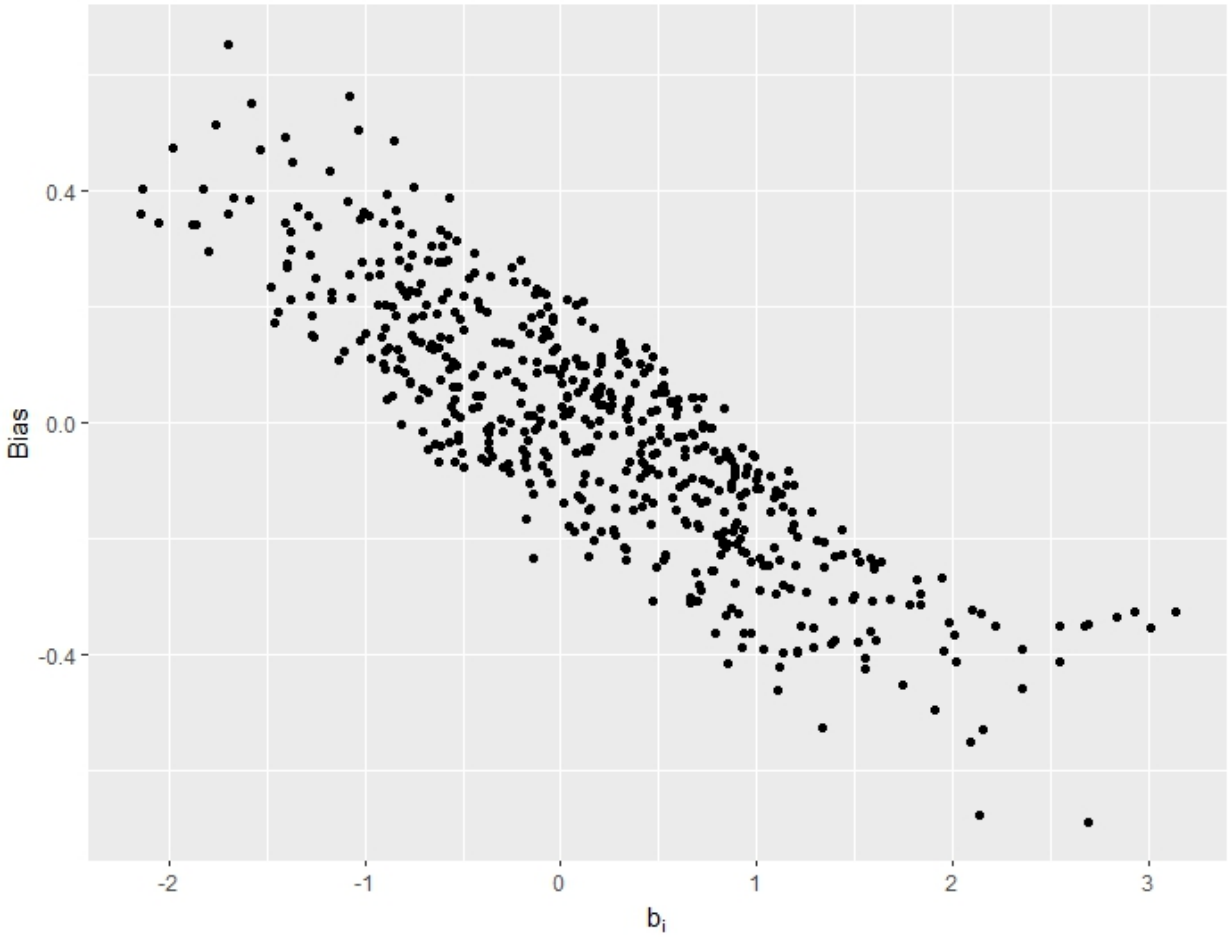

Figure S1. The relationship between response style and bias in trait estimation. Values of $b_i > 0$ reflect DRS or nay-saying, and $b_i < 0$ reflects ARS or yea-saying. The trait depicted here is self-reported Manipulativeness. Bias is calculated as the estimate from the graded response model minus the b_i model. A positive bias indicates the GRM overestimated an individual's degree of Manipulativeness. In general, the GRM overestimates Manipulativeness for those who have a yea-say, and underestimates it for those who nay-say.