Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions

Mirka Henninger & Thorsten Meiser

University of Mannheim

Author Note

This work was supported by the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation (DFG). The authors thank Peter Borkenau and Fritz Ostendorf for providing the data of the standardization sample of the German Big Five questionnaire for the empirical illustration. Furthermore, we would like to thank the anonymous reviewers and Hansjörg Plieninger for helpful comments that substantially helped to improve the manuscript.

Parts of this work have been presented at the *51th Conference of the German Society for Psychology*, Frankfurt am Main, Germany and at the *84th Annual International Meeting of the Psychometric Society*, Santiago de Chile.

Correspondence concerning this article should be addressed to Mirka Henninger, Department of Psychology, University of Mannheim, 68161 Mannheim, Germany.

Email: `m.henninger@uni-mannheim.de`

Abstract

Many approaches in the Item Response Theory (IRT) literature have incorporated response styles to control for potential biases. However, the specific assumptions about response styles are often not made explicit. Having integrated different IRT modeling variants into a superordinate framework, we highlighted assumptions and restrictions of the models (Henninger & Meiser, 2019). In this article, we show that based on the superordinate framework, we can estimate the different models as multidimensional extensions of the Nominal Response Models in standard software environments. Furthermore, we illustrate the differences in estimated parameters, restrictions, and model fit of the IRT variants in a German Big Five standardization sample and show that psychometric models can be used to debias trait estimates. Based on this analysis, we suggest two novel modeling extensions that combine fixed and estimated scoring weights for response style dimensions, or explain discrimination parameters through item attributes. In summary, we highlight possibilities to estimate, apply, and extend psychometric modeling approaches for response styles in order to test hypotheses on response styles through model comparisons.

Keywords: item response theory, response styles, multidimensionality, varying thresholds

Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions

Responses to rating scale items do not only capture the primary trait to be measured, but also response tendencies of the person providing the response (Baumgartner & Steenkamp, 2001). Such response styles are the tendencies of respondents to prefer certain types of categories over others. The tendency of choosing the extreme categories is called *Extreme Response Style* (ERS), of choosing the middle category is called *Mid Response Style* (MRS), and the tendency towards agreeing with the item is called *Aquiescence Response Style* (ARS; Van Vaerenbergh & Thomas, 2013).

Response styles seem to be omnipresent in rating data (e.g., Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel & Carstensen, 2017), consistent across traits (Weijters, Geuens, & Schillewaert, 2010a; Wetzel, 2013), and stable over time (Weijters, Geuens, & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2016). Response styles can influence item responses and therewith bias measurement (see Bolt, Lu, & Kim, 2014; Wetzel & Carstensen, 2017). As an example, Figure 1 shows frequencies of category choices for three exemplary respondents with the same manifest mean across items, but negative, neutral, or positive ERS levels, respectively. Besides, response styles can distort measured relations between variables (Abad, Sorrel, Garcia, & Aluja, 2018; Böckenholt & Meiser, 2017) and comparison between sub-groups, for example in cross-cultural research (Bolt et al., 2014; Rollock & Lui, 2016). Numerous attempts have been proposed in order to control distorting influences of response styles on measurement through questionnaire design and psychometric modeling approaches. As the measurement situation can often not be influenced by the researcher, we focus on psychometric nmodeling approaches to account for response styles in this article.

INSERT FIGURE 1 ABOUT HERE

Psychometric Models for Response Styles

There is a large variety of psychometric modeling approaches accounting for response styles. Here, we examine Divide-by-Total models from Item Response Theory

(IRT) such as the Nominal Response Model, (NRM), or the Partial Credit Model (PCM; see Bock, 1972; Masters, 1982; Takane & de Leeuw, 1987; Thissen & Steinberg, 1986) as they allow us to model response styles in an exploratory as well as confirmatory manner. Within this modeling family, response styles can be incorporated in many different ways: some models have used variations in item thresholds to allow for heterogeneous response scale use, while other models have included additional response style traits. This heterogeneity makes it difficult to identify and assess assumptions that are implicitly made by model constraints. To make such assumptions visible, Henninger and Meiser (2019) integrated the different response style models into a superordinate framework.

In the following, we give a brief summary of this framework and refer to Henninger and Meiser (2019) for more details. In short, response styles can be conceived as person-specific shifts in the thresholds. In consequence, the threshold and category probabilities that describe a response of person n to item i are given by

$$p(X = k|X \in \{k-1, k\}, \theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(\theta_n - b_{ik} + \delta_{nk}\right)}{1 + \exp\left(\theta_n - b_{ik} + \delta_{nk}\right)} \tag{1}$$

and

$$p(X = k|\theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k} b_{ik'} + \sum_{k'=0}^{k} \delta_{nk'}\right)}{\sum_{j=0}^{K} \exp\left(s_j\theta_n - \sum_{k'=0}^{j} b_{ik'} + \sum_{k'=0}^{j} \delta_{nk'}\right)} \tag{2}$$

where k is the response category with $k \in \{0, ..., K\}$, s_k are the scoring weights that are typically fixed to $\boldsymbol{s} = (0, ..., K)$ for ordinal IRT models, θ_n is the respondent's trait parameter, b_{ik} is the item-specific category parameter for item i and category k, and δ_{nk} is a parameter of a person- specific shift in threshold k. Person parameters follow a multivariate normal distribution $[\theta, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. The item-specific category parameter b_{ik} can be decomposed into the item location β_i and threshold τ_{ik} with $\beta_i = (\sum_{k=1}^{K} b_{ik})/K$. For identification, the values of the first category are set to 0 ($s_0\theta_n - b_{i0} + \delta_{n0} \equiv 0$). Further extensions of response style models may include

item-specific discrimination parameters (e.g., $\alpha_i s_k \theta_n - \sum_{k'=0}^{k} b_{ik'} - \alpha_i^{RS} \sum_{k'=0}^{k} \delta_{nk'}$) reflecting the impact of the latent primary trait or response style dimension on single items (see Table 1 and Appendix A in Henninger & Meiser, 2019).

Figure 2 shows how such person-specific threshold shifts can be incorporated in IRT models to reflect response styles. The category probability curves are impacted through the inclusion of response styles into the modeling approach. For example, a respondent with asymmetric threshold shifts has a unique profile of response tendencies that leads to, for example, a decrease in probability for the lowest category (column 2). In contrast, ERS is described by a shift of the outer thresholds towards the item location, thereby widening the interval over which the extreme categories have the modal probability. MRS is described by a shift of the inner thresholds away from the item location, thereby widening the interval over which the middle category is most probable. In consequence, the probability of choosing one of the extreme categories or the middle category increases for a person with a given primary trait level as a function of ERS and MRS, while the probability of choosing one of the intermediate categories decreases (column 3). For positive ARS levels, the threshold separating the middle from the agreement categories is shifted towards the left, increasing the probability of a response in one of the agreement categories (right column).

The formulation of response styles as person-specific threshold shifts δ_{nk} unifies the different psychometric models that have either conceived response styles as variations in thresholds or as additional trait dimensions. To give an example of the latter, in a multidimensional PCM with an additional ERS dimension (e.g., Bolt & Newton, 2011; Wetzel & Carstensen, 2017), the cumulated person-specific thresholds shifts δ_{nk} in Equation 2 are a function of a response style trait θ_n^{ERS} and scoring weights $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ so that $(\theta_n^{ERS}, 0, 0, 0, \theta_n^{ERS})$ reflects the effects of ERS on each category for person n giving a response to an item with five response categories. This example illustrates that we can reparameterize the various response style IRT models in order to describe the composition of δ_{nk}, hence the person-specific shifts in threshold parameters, as additional traits. This reparameterization makes response style

specifications in the different IRT models explicit (for further examples of scoring weights for response style dimensions see Table 1, for the composition of δ_{nk} in different response style models see Tables A1 and A2 in Appendix A in Henninger & Meiser, 2019).

INSERT FIGURE 2 ABOUT HERE

Review of Divide-by-Total Models Accounting for Response Styles

Specifying response styles as person-specific shifts in thresholds highlights which model-implied assumptions have been used in various psychometric approaches. Analyzing these assumptions and restrictions on response styles lead to three groups of models (Henninger & Meiser, 2019).

The first group comprises two models (Wang, Wilson, & Shih, 2006; Wang & Wu, 2011) that account for unknown response styles in the data. The authors see response styles as random noise that can be accounted for by person-specific threshold shift parameters that are independent from each other and from the latent primary traits. As the person-specific threshold shifts are specified as uncorrelated, response styles such as ERS or MRS cannot be captured by the model as they require symmetric threshold shifts across respondents (see Figure 2).

The models in the second group allow for intercorrelations between person-specific thresholds, but still estimate response styles exploratorily. One example of models in this group are mixture distribution models that account for heterogeneity between respondents through assigning them to latent classes with class-specific threshold parameters that can reflect response tendencies (Böckenholt & Meiser, 2017; Eid & Rauber, 2000; Moors, 2003; Rost, 1991). As another example, NRMs have been extended to incorporate an additional response style dimension. The scoring weights $\boldsymbol{s}_k$ of this dimension are estimated freely and can be interpreted post hoc. Another extension was proposed by Bolt et al. (2014) who proposed preference parameters for each category to model the tendency of respondents to prefer certain categories over others. Hence, the second group of models allow researchers to explore the data to find

a common structure of threshold shifts across respondents.

The third group of models use a priori specifications of response styles. For example, Jin and Wang (2014) assumed that response styles pull apart or push together item thresholds. They introduced a person-specific weight parameter to reflect this dispersion. Other approaches added response style dimensions to a PCM. The scoring weights $\boldsymbol{s}_k$ of response style dimensions are fixed a priori, for example to incorporate ERS, MRS, and ARS traits into the model (see column 3 & 4 in Figure 2 and Table 1). In consequence, correlations between response style and primary trait dimensions can be examined (Bolt & Newton, 2011; Tutz, Schauberger, & Berger, 2018; Wetzel & Carstensen, 2017). Falk and Cai (2016) added item-specific discrimination parameters to describe the impact of response style dimensions on items as a further extension.

Implications of the Integrated Framework and Overview

All response style models from the Divide-by-Total framework can be written in the notation of multidimensional NRMs (Thissen & Steinberg, 1986). Through this notation, we can derive scoring weights $\boldsymbol{s}_k$ that in turn allow us to estimate the different models as multidimensional extensions of the NRM in standard software environments such as Mplus (Muthén & Muthén, 2012) or the statistical environment *R* (R Core Team, 2019). We have collected scoring weights for the response style models in Table 1 and provide a short introduction on model estimation in the next section.

Furthermore, knowing about the assumptions and restrictions of the response style models allows us to test these assumptions in empirical data. For example, we can examine whether response styles are unsystematic noise in rating data (see Wang et al., 2006), whether there are systematic response style effects across respondents (see Bolt & Johnson, 2009; Bolt et al., 2014), or whether there are substantial latent correlations between primary trait and response style dimensions (see Falk & Cai, 2016; Wetzel & Carstensen, 2017). To demonstrate such comparisons, in the remainder of this article we illustrate the estimation of these models with a Big Five standardization sample, give an overview on model specification, highlight the parameters that are estimated in

each modeling approach, and interpret model fit.

In addition, we can use the superordinate framework to derive novel extensions to the existing models. In this vein, we propose two novel model variants that extend existing IRT models for response styles. The first proposition combines approaches with fixed and estimated scoring weights. In the second proposition, we combine methods from explanatory IRT modeling (De Boeck & Wilson, 2004; Embretson, 1999) and response style modeling and specify discrimination parameters as functions of item attributes. Both models fit in and extend the model structure, and open up new possibilities to improve measurement of traits and analyses of response styles.

Model Implementation in Standard Software

Subsuming the different Divide-by-Total modeling extensions under the superordinate framework (Equations 1 and 2) allows us to implement the models as multidimensional extensions of NRMs (Bock, 1972; Takane & de Leeuw, 1987). As it is not immediately obvious how threshold shifts translate into scoring weights—in particular for models with varying thresholds (e.g., Wang et al., 2006), or for models with category preferences summing up to zero (Bolt et al., 2014)—expressing response style models as multidimensional NRMs allows us to identify the scoring weights that we can use for estimation in standard statistical software (see Henninger & Meiser, 2019). We summarize scoring weights for trait, random thresholds, exploratory response styles, category preferences, and response styles ERS, MRS, and ARS for an item with $K = 4$ thresholds and $K + 1 = 5$ categories in Table 1.

Standard statistical software programs such as Mplus (Muthén & Muthén, 2012, see also Huggins-Manley & Algina, 2015) or the *R* (R Core Team, 2019) packages *TAM* (Kiefer, Robitzsch, & Wu, 2017) or *mirt* (Chalmers, 2012) have built-in procedures to estimate multidimensional IRT models. Standard software programs implement procedures that allow us to specify whether scoring weights of each item and category for each latent dimension should be estimated, constrained, or fixed to a specific value. For example, we can set up a multidimensional PCM with fixed scoring weights for trait

and response style dimensions through specifying that each item relates to both, the primary trait and the response style dimensions through the scoring weights from Table 1 (e.g., $\boldsymbol{s} = (0, 1, 2, 3, 4)$ for the trait and $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ for ERS).

——————— INSERT TABLE 1 ABOUT HERE ———————

We give an example of such a within-item multidimensionality scoring procedure for estimation in the *R* package *TAM* in Appendix A with scoring weights for two primary trait dimensions with four items each (2 of which are reversed coded) and response styles ERS and MRS that load on all eight items. Hence, response styles ERS and MRS are constrained to be equal across primary dimensions. In addition to the example in the Appendix, we provide code and instructions on how to implement response style models (PCM ignoring response styles as well as models by Bolt & Johnson, 2009; Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Wang et al., 2006; Wang & Wu, 2011; Wetzel & Carstensen, 2017) in *TAM* based on a simulated dataset with the same data structure as the data in the following empirical analysis on Github[1].

Model Comparison Using Empirical Data

The integration of response style modeling approaches into one superordinate framework (Henninger & Meiser, 2019) allows us to estimate the models within one software environment as multidimensional NRMs, and to compare models with different assumptions on response styles. In consequence, we can examine whether response styles are best reflect by a notion of random noise (e.g., Wang et al., 2006), or whether a common structure of threshold shifts exists in the data (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011). In addition, we use this model illustration to feature the different specification and estimated parameters of the response style models for applied research and to demonstrate how primary trait estimates can be corrected for response style models through psychometric approaches.

We analyzed a non-clinical standardization sample of a German Big Five inventory

[1] `https://github.com/mirka-henninger/FitResponseStyles`

by Borkenau and Ostendorf (2008). In this sample, 11,724 respondents answered a Big Five questionnaire, wherein each scale consists of twelve items on a 5-point rating scale, hence 60 items in total. As baseline models, we fit a PCM and a generalized PCM with discrimination parameters, both ignoring response styles, to the Big Five data. We chose the PCM and generalized PCM as a special case of the NRM with fixed scoring weights for the Big Five dimensions, as the (g)PCM reflects the ordinal structure of the response categories, while a NRM with estimated scoring weights is rather suited to model responses to nominal categories (Thissen & Steinberg, 1986).

We selected a sample of the Divide-by-Total response style models. First, we chose models with continuous parameterization of response styles, and hence excluded mixture IRT model and latent class factor models (Moors, 2003; Morren, Gelissen, & Vermunt, 2011; Rost, 1991). Furthermore, we chose models with the ability to account for several response tendencies, for example modeling random thresholds, several response style dimensions exploratorily, category preferences, or pre-specified response styles such as ERS, MRS, and ARS. This selection excluded the model by Jin and Wang (2014) and Tutz et al. (2018) because they solely incorporate ERS/MRS. Our selection therefore comprised six response style models: a random threshold model (Wang et al., 2006), a generalized random threshold model with item discrimination parameters (adapted from Wang & Wu, 2011), a multidimensional NRM with freely estimated scoring weights for response styles (Bolt & Johnson, 2009), a model with category preferences parameters for response styles (Bolt et al., 2014), a multidimensional PCM with fixed scoring weights for response styles (Bolt & Newton, 2011; Wetzel & Carstensen, 2017) and a generalized multidimensional PCM with item-specific discrimination parameters (Falk & Cai, 2016).

Response styles were modeled across all 60 Big Five items and with the same scoring for reversed and non-reversed items (see Table 1 and Wetzel & Carstensen, 2017, for a discussion on using the same, separate, or additional items for the response style dimension). All models were estimated using R (R Core Team, 2019) with the package *TAM* (Test Analysis Modules, Kiefer et al., 2017). Within *TAM*, we used the

marginal maximum likelihood method to estimate multidimensional IRT models with estimated or fixed scoring weights and discrimination parameters. For high dimensional models, TAM offers a quasi Monte-Carlo integration procedure (Pan & Thompson, 2007) that prevents the time intensive numerical integration.

Model Specification

For all models, we estimated the Big Five trait dimensions Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness using fixed scoring weights $\boldsymbol{s}^{BigFive} = (0, 1, 2, 3, 4)$ or $\boldsymbol{s}^{BigFiveReversed} = (4, 3, 2, 1, 0)$ for reversed coded items and allowed the Big Five dimensions to correlate with each other. The PCM had 255 parameters (240 fixed item-threshold parameters, 5 latent trait variances for the Big Five dimensions, and 10 latent covariances between dimensions with $\boldsymbol{\theta}^{BigFive} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. The generalized PCM had 310 parameters (240 fixed item-threshold parameters, 60 discrimination parameters, 5 latent trait variances for the Big Five dimensions were fixed to 1, and 10 latent covariances between dimensions were estimated with $\boldsymbol{\theta}^{BigFive} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$.

Scoring weights to specify the random threshold model (Wang et al., 2006) are presented in Table 1. Here, 259 parameters where estimated (240 item-threshold parameters, 5 Big Five variances, 4 threshold variances, and 10 latent covariances between Big Five dimensions with $[\boldsymbol{\theta}^{BigFive}, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where covariances were fixed to 0 between Big Five dimensions and thresholds, as well as between random thresholds). The same scoring weights were used for the generalized random threshold model (adapted from Wang & Wu, 2011), in which we estimated 60 additional discrimination parameters for the Big Five dimensions, and 60 discrimination parameters for the random threshold dimensions. Hence, 370 parameters were estimated (240 item-threshold parameters, 120 discrimination parameters, 5 Big Five variances, and 4 threshold variances were fixed to 1 with $[\boldsymbol{\theta}^{BigFive}, \delta_1, ..., \delta_K]$ $\sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where, as before, 10 latent covariances between Big Five dimensions were estimated, while covariances were fixed to 0 between Big Five dimensions and

thresholds, as well as between random thresholds)[2].

For the multidimensional NRM (Bolt & Johnson, 2009), scoring weights for the Big Five dimensions were fixed, while scoring weights for three response style dimensions were estimated. This results in 270 estimated parameters (240 item-threshold parameters, 15 scoring weight parameters, one for each of the five categories relating to the three response style traits, 5 Big Five variances, and 10 latent covariances between Big Five dimensions, were estimated, and 3 response style trait variances were fixed to 1 with $[\boldsymbol{\theta}^{BigFive}, \theta^{RS1}, \theta^{RS2}, \theta^{RS3}] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where covariances were fixed to 0 between Big Five and response style dimensions, as well as between response style dimensions).

For the model with category preference parameters for response styles (Bolt et al., 2014), scoring weights for the Big Five and the category preference dimensions were fixed (see Table 1). This results in 285 estimated parameters (240 item-threshold parameters, 5 Big Five variances, 4 category preference variances and 36 latent covariances with $[\boldsymbol{\theta}^{BigFive}, \theta_1, ..., \theta_4] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$; the last category preference parameter can be derived from the others as across categories they sum to 0).

For the multidimensional PCM with response styles ERS, MRS, and ARS (Bolt & Newton, 2011; Wetzel & Carstensen, 2017), scoring weights for the Big Five and response style dimensions were fixed (see Table 1). This results in 276 estimated parameters (240 item-threshold parameters, 5 Big Five variances, 3 response style variances and 28 latent covariances with $[\boldsymbol{\theta}^{BigFive}, \theta^{ERS}, \theta^{MRS}, \theta^{ARS}] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$).

The generalized multidimensional PCM with response styles ERS, MRS, and ARS

---

[2] Please note that this is not the original model proposed by Wang and Wu (2011), but an extension thereof. Wang and Wu (2011) assumed that item-specific discrimination parameters are equal across all latent dimensions, that is the latent trait and the K random thresholds. The assumption that discrimination parameters are equal for the traits and random thresholds seems not plausible, however, and hinders the interpretation of discrimination parameters since it is unclear whether they reflect traits or response styles. Therefore, we extended the model for a new set of discrimination parameter that differentiates between discrimination parameters related to the trait and random thresholds. In this analysis, we restricted discrimination parameters to be equal between random threshold dimensions.

(Falk & Cai, 2016) used fixed scoring weights for the Big Five and response style dimensions, but estimated discrimination parameters for the Big Five traits and response styles. This results in 449 estimated parameters (240 item-threshold parameters, 181 discrimination parameters, whereof 60 for Big Five traits, 60 for ERS, 60 for MRS, and 1 for all ARS indicators, see also Maydeu-Olivares & Coffman, 2006, 5 Big Five variances and 3 response style variances were fixed to 1 and 28 latent covariances with $[\boldsymbol{\theta}^{BigFive}, \theta^{ERS}, \theta^{MRS}, \theta^{ARS}] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ were estimated).

Model Fit

Table 2 gives an overview of the estimated parameters as well as model fit indices for the IRT models in the application to the German Big Five standardization sample. We evaluated absolute model fit in terms of the Standardized Generalized Dimensionality Discrepancy Measure (SGDDM; Levy, Xu, Yel, & Svetina, 2015). This measure can be interpreted in the metric of a correlation where values close to 0 indicate good fit and little local dependence. According to SGDDM all models display values close to 0 and we find no substantial differences in absolute model fit. Furthermore, we report the Log-Likelihood and Bayesian Information Criteria (BIC; Schwarz, 1978). For model comparisons, we used Likelihood-Ratio tests with the PCM as a reference model to examine the increase in model fit when response styles are accounted for. In case of the generalized response style models by Wang and Wu (2011) and Falk and Cai (2016), we used the generalized PCM as a reference. We base our model comparison (e.g., the rank order in Table 2) on BIC due to its ease of interpretation and penalty for additional model parameters, but also extend model comparisons by χ^2 tests between response style models in the following discussion, where applicable.

Overall, accounting for response styles clearly led to better model fit (all $\chi^2 \geq 30,182$, $p < .001$). Based on BIC, it appears that we can find a common structure of threshold shifts across respondents (Bolt & Johnson, 2009; Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017) as models specifying response styles as random noise had a worse model fit (Wang et al., 2006; Wang & Wu,

2011). Similarly, we find that allowing for latent covariances between traits and response style dimensions seems to be a sensible approach in this dataset (Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017, with an exception of the multidimensional NRM by Bolt & Johnson, 2009, and Bolt et al., 2014). Finally, we see that latent dimensions impact items differently as approaches with item-specific discrimination parameters showed an improved model fit (Random Threshold Model vs. generalized Random Threshold Model: $\chi^2(111) = 9,778$, $p < .001$; multidimensional PCM vs. generalized multidimensional PCM: $\chi^2(173) = 14,623$, $p < .001$).

——————— INSERT TABLE 2 ABOUT HERE ———————

All together, the model that fit the data best compared to the other models was the generalized multidimensional PCM with ERS, MRS, and ARS response style dimensions and discrimination parameters for trait and response styles (Falk & Cai, 2016). Table 3 shows the estimated variance-covariance matrix of the model. We can see that MRS and ARS were moderately related, and that the Agreeableness dimension shows negative correlations with ERS and ARS.

——————— INSERT TABLE 3 ABOUT HERE ———————

Furthermore, the superior model fit due to estimated discrimination parameters suggests that items were differentially impacted by the latent dimensions, Big Five dimensions as well as response style dimensions. Overall, the ERS dimension had a larger impact ($\bar{\alpha}^{ERS} = 1.10$) than the MRS ($\bar{\alpha}^{MRS} = 0.60$), or ARS dimensions ($\alpha^{ARS} = 0.18$; all latent trait variances were fixed to 1). Figure 3 illustrates the impact of the ERS dimension on two items, one with the lowest ($\alpha^{ERS}_{min} = 0.53$; upper panel) and the other with the highest discrimination ($\alpha^{ERS}_{max} = 1.62$; lower panel). We can see that threshold and category probability curves of the item with low discrimination were nearly unaffected by the latent ERS dimension (probabilities were largely independent of ERS trait levels). In contrast, threshold and category probability curves were noticeably affected when discrimination was high (probabilities were largely dependent on ERS trait levels). Hence, accounting for differential influence of the response style dimensions on items seems to play a substantial role in this dataset.

INSERT FIGURE 3 ABOUT HERE

Figure 4 shows how including response styles into psychometric models can debias the primary trait estimates—here for the Neuroticism trait. The upper panel depicts response category choices for different ERS levels (lower and upper 10% and intermediate levels) based on the generalized multidimensional PCM by Falk and Cai (2016). For the lowest 10%, the extreme categories are rarely chosen, while the opposite occurs for respondents with highly positive ERS levels. The lower panel shows trait estimates of the Neuroticism dimension under a generalized PCM ignoring response styles, and the model by Falk and Cai (2016) with additional ERS, MRS, and ARS dimensions for varying levels of ERS. For low Neuroticism and high ERS trait levels, the response style model performs an upwards correction of Neuroticism estimates as the "strongly disagree" category is chosen inappropriately often; however, a downwards correction occurs for high Neuroticism trait levels as the "strongly agree" category is chosen inappropriately often. The opposite correction occurs for respondents with low ERS trait levels that tend to avoid the extreme categories. Hence, the scatterplot in the lower panel of Figure 4 demonstrates how a psychometric model can correct for biasing effects of response styles through allowing for a priori specified shifts in threshold parameters for different respondents.

INSERT FIGURE 4 ABOUT HERE

To conclude, in the Big Five standardization sample, a clear advantage of models specifying response styles a priori and therefore allowing for covariances between traits, between response styles, and between traits and response styles (models by Bolt et al., 2014; Falk & Cai, 2016; Wetzel & Carstensen, 2017) was found in the data. Besides the increased model fit in this dataset, IRT variants with a priori specified response styles have a straight-forward interpretation of response style dimensions and the relation between latent dimensions can be explored through the variance-covariance matrix $\mathbf{\Sigma}$. In addition, an advantage of models using item-specific discrimination parameters emerged. Such or similar comparisons between response style models can be useful tools to test specific assumptions on response styles in order to build a coherent theoretical

framework for response styles.

New Model Extensions

In the Big Five standardization sample, we found a superiority of models specifying response styles a priori and allowing for differential influences of the response style dimensions on single items. However, both model specifications come at a price. First, specifying response style a priori implies strong assumptions on response style specifications, namely symmetric threshold shifts for ERS and MRS around the item location ($\theta_n^{ERS} = -\delta_{n1} = \delta_{n4}$; $\theta_n^{MRS} = \delta_{n2} = -\delta_{n3}$ for an item with 4 thresholds). For ARS, scoring weights $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ stand for a shift in the third threshold, while the threshold probability of the highest thresholds stays constant (see Figure 2 and Appendix B in Henninger & Meiser, 2019). Second, when including discrimination parameters for response style dimensions (Falk & Cai, 2016), the model becomes highly flexible through allowing the dimensions to have differential influences on the items. However, the number of estimated parameters increases tremendously, especially when the number of latent (response style) dimensions is large. Furthermore, even though we now allow response style dimensions to impact certain items to a larger extent than others, it remains unknown why this is the case.

In this section, we propose two new model extensions that address these challenges and fill in gaps in the model structure. The first model lifts equality constraints on scoring weights (and therewith threshold shifts) in multidimensional PCMs for ERS, MRS, and ARS. It is more flexible than fixing the scoring weights a priori for all categories (e.g., Wetzel & Carstensen, 2017), but defines the type of response style in contrast to a multidimensional NRM with estimated scoring weights (Bolt & Johnson, 2009). The second model combines methods from explanatory IRT models (De Boeck & Wilson, 2004; Embretson, 1999) with response style modeling. It defines discrimination parameters as a function of item attributes, and therefore is more restrictive and parsimonious than the model by Falk and Cai (2016), but has a higher flexibility than a multidimensional PCM (e.g., Wetzel & Carstensen, 2017). In addition,

it allows us to identify item attributes that enhance the influence of response style dimensions on item responses.

After briefly introducing the two new model variants, we use the Big Five standardization sample to fit examples of the two models to extend and complete the model structure and the illustration with empirical data of the previous section.

Combining Fixed and Estimated Scoring Weights

In order to test whether the effect of ERS is stronger for the agreement than the disagreement categories or vice versa, whether MRS not only affects the middle, but also the intermediate categories, or whether the two agreement categories are differentially affected by ARS, we propose a new IRT model variant lifting the equality constraint on category scoring weights. Instead of estimating the scoring weights freely (Bolt & Johnson, 2009), or fixing them a priori (Bolt & Newton, 2011; Wetzel & Carstensen, 2017), we defined a more parsimonious, or flexible approach, respectively. For this purpose, we specified new scoring weights that are partly fixed and partly estimated to test whether response style traits affect specific categories differently within items. The resulting scoring weights for response style traits for an item with 5 response categories can be specified as:

$$\begin{aligned} \boldsymbol{s}^{ERS} &= (1, 0, 0, 0, \lambda^{ERS}) \\ \boldsymbol{s}^{MRS} &= (0, \lambda^{MRS}, 1, \lambda^{MRS}, 0) \\ \boldsymbol{s}^{ARS} &= (0, 0, 0, 1, \lambda^{ARS}). \end{aligned} \tag{3}$$

The additional, estimated scoring weight parameter λ that is equal across participants and items reflects the assumption that effects of response styles on categories may not be the same for all categories, but proportional between categories within items. For example for ERS, the extreme categories are not affected equally, but we can test whether the highest category is affected more strongly than the lower category. When $\lambda^{ERS} > 1$, θ_n^{ERS} affects the highest agreement category more strongly than the lowest disagreement category and vice versa for $\lambda^{ERS} < 1$. When $\lambda^{MRS} > 0$ not only the

probability for the middle, but also the probability for intermediate categories increases for positive levels of θ_n^{MRS3}. $\lambda^{ARS} > 1$ implies that θ_n^{ARS} influences the highest threshold and hence increases probability of choosing the highest category more strongly. Therewith, λ^{ARS} makes the assumption that ARS affects only certain threshold shifts testable (see the rightmost column in Figure 2 for shifts in threshold when $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$).

Modeling Discrimination Parameters Through Item Attributes

Response styles may have stronger or weaker influences on item responses depending on item attributes, such as item complexity or item position. To propose a parsimonious model where the differential influence of the response style dimensions on items is captured by item-specific discrimination parameters α_{id}, they can be defined as functions of item attributes. Such attributes can be contextual influences, such as the number of response options, item wording, ambiguity, complexity, negation, reversal, or position effects, for instance:

$$\alpha_{id}^{*} = f(Complexity_i, Negation_i, Position_i). \tag{4}$$

The function f can be a linear parameter combination of item attributes; but also other kinds of function may apply. Hence, this model can be regarded as an explanatory IRT approach for discrimination parameters to investigate the impact of response style dimensions on specific item types (see Cho, De Boeck, Embretson, & Rabe-Hesketh, 2014; De Boeck & Wilson, 2004; Embretson, 1999, for explanatory IRT approaches).

Fit of New Model Extensions to the Big Five Standardization Sample

We fit two exemplary specifications of the new modeling variants to the Big Five standardization sample. Of course, the approaches presented here serve as a guidance

[3] Alternatively, $\boldsymbol{s}^{MRS} = (0, 0, 1, 1 - \lambda^{MRS}, 0)$ allows one to test whether threshold shifts are symmetric around the middle category; when $\lambda^{MRS} = 1$, a symmetry of threshold shifts in line with column 3 in Figure 2 is given.

for applications and can be specified for any other latent dimension or adapted for other types of attributes or alternative explanatory approaches. For the model combining fixed and estimated scoring weights, we used the response style dimension ARS as an example. In particular for ARS, different scoring weights have been proposed in psychometric approaches (see e.g., Billiet & McClendon, 2000; Falk & Cai, 2016; Plieninger & Heck, 2018; Weijters et al., 2010b; Wetzel & Carstensen, 2017, and Table 1). By specifying scoring weights for the ARS dimension that are partly fixed and partly estimated such as $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, \lambda^{ARS})$, we can test whether and to which magnitude the ARS dimension affects the upper threshold. All other parameter were specified as in the model of Wetzel and Carstensen (2017) in the illustration section above.

To specify a model with constrained discrimination parameters of response style dimensions ERS and MRS, we used three types of item attributes to define the restrictions: item negation, complexity and position (see Table 4 and Table B1 in Appendix B). Items received the value 1 when they were negated (e.g., contained "not", "not a", "never") and 0 otherwise; items were coded 1 on Complexity if the item content included more than one line of thought (i.e. double-bind items, e.g., "I am quite good at organizing my time for myself so that I can finish my affairs on time."). Please note that item responses in the 60 item version of the Big Five standardization sample used in for analyses herein were collected with a 240 item measure. We used the position of items from the 240 item instrument, so item received the value 1 when they occurred in the last half of the 240 item instrument, and 0 otherwise. We used a linear model with a fixed-links approach (e.g., Schweizer, 2008; Zeller, Reiß, & Schweizer, 2017) to decompose α_{id} into elementary parameters. We defined these elementary parameters to be an intercept ($\alpha_{Intercept}$) and effects for each of the item attributes ($\alpha_{Negation}$, $\alpha_{Complexity}$, $\alpha_{Position}$; see Table 4). Hence, we estimated four discrimination parameters for each of the response style dimensions ERS and MRS while fixing their latent variances to one (analogous to Falk & Cai, 2016). Discrimination parameters were fixed to one for the Big Five dimension and ARS dimension to facilitate interpretation (all other parameters were specified as in the multidimensional PCM in

the illustration section). The model therewith allows us to examine the moderating role of item attributes on response style effects in a focused way.

INSERT TABLE 4 ABOUT HERE

Table 5 extends the overview of estimated parameters and information criteria of the response style models (Henninger & Meiser, 2019) by the two exemplary modeling extensions. The model combining fixed and estimated scoring weights for the ARS trait fits the data better than its restricted variant (Wetzel & Carstensen, 2017, $\chi^2(1) = 141$, $p < .001$). The scoring weights are $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, \lambda^{ARS})$, with $\lambda^{ARS} = 1.36, SE < 0.01$. This indicates that for the ARS trait, not only the third threshold is shifted by θ_n^{ARS}, but also the upper threshold is shifted by $0.36 \cdot \theta_n^{ARS}$. Stated differently, this response style model variant shows that the threshold probability between the two agreement categories is affected by the ARS trait, but to a lower degree than the threshold between the middle and the first agreement category.

INSERT TABLE 5 ABOUT HERE

In the new model variant imposing constraints based on item attributes on discrimination parameters of the ERS and MRS latent traits, four item discrimination parameters were estimated for each of the two response style dimensions (see Table 6). Hence, the new restrictions reduced the number of discrimination parameters from 60 to four for each of the two response style dimensions (i.e. reducing 112 parameters in total). Unsurprisingly, the restricted model has a worse fit than the generalized multidimensional PCM (Falk & Cai, 2016, $\chi^2(167) = 13,893$, $p < .001$), as we would not assume that the reduction in estimated parameters and model flexibility goes unnoticed. However, there is still a substantive increase in model fit compared to the multidimensional PCM with response styles based on a 1-parameter model with item-invariant discrimination parameters (Wetzel & Carstensen, 2017, $\chi^2(6) = 731$, $p < .001$) which speaks in favor of the utility of using information on item attributes for parameter estimation[4].

[4] As a competitor model, we fit an alternative approach with four discrimination parameters for the ERS as well as the MRS dimension. As in the proposed model, one discrimination parameter reflected

Table 6 shows the four elementary parameters for the discriminatory weights in the ERS and MRS dimension. The intercept reflects the discrimination of non-negated, non-complex items that appeared in the first half of the 240 item Big Five questionnaire. We see that averaged across these items, ERS has a larger influence on item responses than MRS. This difference in $\alpha_{Intercept}$ directly reflects the difference in variance between ERS and MRS dimensions that is typically found in empirical data (e.g., Plieninger & Heck, 2018; Wang et al., 2006; Wetzel, Böhnke, & Rose, 2016). We set the significance level to $\alpha = .001$ due to the multiple tests and large sample size. Results indicate that negated items and items appearing in the second half of the questionnaire increased the influence of ERS and MRS on item responses. Complex items, in contrast, decreased the influence of ERS, while there was no effect on the strength of MRS. The effect of item position on discrimination parameters of the ERS dimension is particularly large ($b = 0.13, t = 20.49$), indicating that items are more strongly influenced by ERS when they appear later in the questionnaire (see also Figure 3 for an illustration of the influence of ERS on two items with different discrimination parameters). To summarize, the results indicate that the impact of response styles on item responses can be defined as a function of item attributes and assessed as such using psychometric modeling approaches for response styles.

INSERT TABLE 6 ABOUT HERE

Discussion

A variety of IRT model extensions accounting for response styles can be subsumed under the superordinate framework of shifting thresholds (see Henninger & Meiser, 2019). Based on the framework, the models can be structured in three groups: models with unique individual profiles of response tendencies (Wang et al., 2006; Wang & Wu,

---

the intercept, while all others were randomly assigned to the 60 Big Five items. In consequence, item characteristics could not have any systematic influence on discrimination of the latent traits. The competitor model fit the data worse ($\Delta BIC = -155$) than the model incorporating item characteristics further suggesting that variations of item attributes systematically affect the impact of response styles on item responses.

2011), models exploring common response styles across respondents in the data (Böckenholt & Meiser, 2017; Bolt & Johnson, 2009; Bolt et al., 2014; Moors, 2003; Rost, 1991), and models specifying response styles a priori (Bolt & Newton, 2011; Falk & Cai, 2016; Jin & Wang, 2014; Morren et al., 2011; Wetzel & Carstensen, 2017).

As all modeling extensions can be written as multidimensional NRMs, we can derive scoring weights for each of the models. These scoring weights can in turn be used to estimate the models in standard software, for example in Mplus (Muthén & Muthén, 2012, see also Huggins-Manley & Algina, 2015) or the statistical programming environment *R* (R Core Team, 2019) using packages *mirt* (Chalmers, 2012) or *TAM* (Kiefer et al., 2017).

We can use psychometric models for response styles to improve primary trait estimation in psychological assessment. In the model illustration, we showed how trait estimates can be debiased through response style modeling, and also how we can use model comparison to test assumptions on response styles. We found a superiority of models that accounted for response styles, specified response styles a priori and therewith were able to estimate relations between latent dimensions, and of models that allowed for a differential impact of latent dimensions on the items.

Building on these results, we proposed two novel types of model extensions that add to the structure of response style models (see Table 1 in Henninger & Meiser, 2019). The first modeling extension combines approaches fixing and estimating scoring weights for response style dimensions. This approach allows us to define the type of response style, while at the same time testing specific symmetry constraints or a priori fixations. We used this semi-exploratory approach to examine the magnitude to which ARS affects the highest threshold by using the Big Five standardization sample (Borkenau & Ostendorf, 2008). The second approach combines explanatory IRT (De Boeck & Wilson, 2004; Embretson, 1999) with response style modeling. We defined item-specific discrimination parameters as a linear function of item attributes. Therewith, we were able to investigate the influence of negation, complexity, and position on the impact of response style dimensions on item responses. We find that

ERS and MRS have a larger impact on negated items, and on items that appeared in the second half of a 240 item questionnaire. Surprisingly, complex items seem to reduce the impact of ERS. As a posthoc interpretation we suggest that complex items may be associated with increased cognitive effort that reduces response style influence. At the same time, we would like to emphasize that these results may be specific for the Big Five dataset used in the analysis. Therefore, the novel models serve as an illustration how fixed and estimated scoring weights can be combined, and how explanatory IRT can be integrated in response style modeling.

Disentangling Primary Trait and Response Style Dimensions in Estimation

The different psychometric approaches for response styles that are presented and discussed in this article and in Henninger and Meiser (2019) impose various assumptions and restrictions on response styles, and this in turn sets specific requirements for the data structure. For instance, few reversed-coded items or a small variance in item difficulty parameters (all items are similarly easy or difficult) may hinder model estimation, resulting in large standard errors and convergence problems (see Johnson & Bolt, 2010). Hence, when designing the questionnaire, one should pay careful attention to include items with different difficulties and reversed-coded items (see also Billiet & McClendon, 2000; Plieninger, 2017). Model estimation can further be facilitated by using additional information for response style dimensions, such as extraneous items (Wetzel & Carstensen, 2017) or anchoring vignettes (Bolt et al., 2014, but see also von Davier, Shin, Khorramdel, & Stankov, 2018). When convergence problems persist, one may consider the use of uncorrelated factors to facilitate separation of primary trait and response style dimensions (Johnson & Bolt, 2010).

As Johnson and Bolt (2010) noted exploratory response style models are particularly vulnerable with regards to empirical identification problems (e.g., models by Bolt & Johnson, 2009; Bolt et al., 2014; Moors, 2003). Also our new model proposition that combines fixed and estimated scoring weights may hinder estimation for certain data structures. For example, combining an ERS dimension with scoring

weights $\boldsymbol{s}_k^{ERS} = (1, 0, 0, 0, \lambda^{ERS})$ with an ARS dimension using $\boldsymbol{s}_k^{ARS} = (0, 0, 0, 1, 1)$ may lead to a confound when no reversed-coded items are present to separate ARS from ERS (see also Billiet & McClendon, 2000; Plieninger, 2017). Johnson and Bolt (2010, p. 97ff, 102, and 109) and Falk and Cai (2016, p. 333-334 and Appendix A) provide further details on statistical and empirical identification strategies.

Alternative Approaches to Account for Response Styles

This article focused on multidimensional extensions of Divide-by-Total models to account for response styles, but there are also other psychometric models based on the graded response model (Samejima, 1969), sequential (Tutz, 1997) or step models (Verhelst, Glas, & De Vries, 1997) and IRTree models (Böckenholt, 2012; De Boeck & Partchev, 2012), or design-based approaches that we would like to mention briefly.

Based on the graded response model, threshold models accounting for response styles via person-specific location and scale parameters (Rossi, Gilula, & Allenby, 2001), varying threshold distances between respondents (Johnson, 2003), or a two-step decision process (Thissen-Roe & Thissen, 2013) have been proposed. Other models are based on the idea of unfolding models to account for ERS (Javaras & Ripley, 2007), or testlet approaches (Bradlow, Wainer, & Wang, 1999), where a general ERS dimension and specific trait dimensions are modeled (De Jong, Steenkamp, Fox, & Baumgartner, 2008). More recent models specify covariates to disentangle trait and response style in a one-item, adjacent categories model (Tutz & Berger, 2016), or adapted the differential discrimination model (Ferrando, 2014) to ordinal responses (Lubbe & Schuster, 2017). Recently, Böckenholt (2012) and De Boeck and Partchev (2012) proposed IRTree approaches that define responses to rating scale items as a sequence of multiple processes (see also Jeon & De Boeck for multiple dimensions and inclusion of covariates; Khorramdel & von Davier, 2014, for a multi-scale extension; Meiser, Plieninger, & Henninger, 2019, for an extension to ordinal judgment processes; Plieninger & Heck, 2018, for an extension for ARS; Plieninger & Meiser, 2014, for a test of validity; Zettler, Lang, Hülsheger, & Hilbig,2016, for an application).

At the same time, attempts to control for response style during measurement have been made. For example, situational factors such as respondents' motivation or cognitive load (Cabooter, 2010), or features of the questionnaire format such as the number of categories, response option labels, reverse-coded or negated items (Weijters et al., 2010b) may reduce response biases. In the multidimensional forced-choice format respondents rank groups (e.g., triplets) of items depending on how well they describe their behavior (see Brown & Maydeu-Olivares, 2013, for a Thurstonian IRT approach to handle ipsative data arising from this format). Instead of giving rating responses, some response format ask respondents to sort items into the response categories (e.g., McKeown & Thomas, 1988; Thurstone, 1928). Current research focuses on the power to reduce response style effects by these and other response formats (see for example Böckenholt, 2017; Plieninger, Henninger, & Meiser, 2019, for experimental investigations).

Directions for Future Research

Response styles should not only be seen as nuisance variables that have to be controlled for, but analyzed as part of a psychologically meaningful response process. To understand the nature of response styles, we must investigate situational and interindividual factors. Hamilton (1968) and Van Vaerenbergh and Thomas (2013) summarized evidence for relationships between response styles and personality variables; however, most results are mixed. Sensible starting points to further increase knowledge on response tendencies themselves are, first, integrating response styles in their nomological net by investigating their relation to personality covariates. These covariates, however, should be measured by response-style-free methods, such as the multidimensional forced-choice method (Brown & Maydeu-Olivares, 2011), the drag-and-drop format (Böckenholt, 2017) or implicit methods (Schmukle, Back, & Egloff, 2008) to avoid confounding effects of response styles. Second, one should examine response processes that moderate the use of response styles in a given questionnaire item. For example, one could analyze how response times moderate response style

effects on category choice. Such investigations would inform us about response styles themselves, their relation to item content, and processes underlying item responses.

The advancement of existing models is a further route for future research. For instance, the random threshold model by Wang and colleagues (Wang et al., 2006; Wang & Wu, 2011) is a promising candidate for modeling response styles as it allows researchers to model heterogeneity towards any response category with little a priori assumptions. This is particularly important when comparing different subgroups with unknown response styles, as might, for example, be the case in cross-cultural research. However, as demonstrated in the application, the model is likely to be violated in empirical data due to the independence restriction on the variance-covariance matrix. Furthermore, it is not possible to interpret person-specific threshold effects in terms of ERS or MRS, because then response styles induce a non-diagonal variance-covariance matrix of person effects. Therefore, more flexibility in the random threshold model concerning its identification constraints is desirable, allowing to estimate the variance-covariance matrix.

The generalized multidimensional PCM with constraints on discrimination parameters of the response style dimensions that we proposed as a model extension also opens up routes for future research. In this approach, we have modeled discrimination parameters of response style dimensions as a function of item attributes, such as position, negation, or complexity. Herein, we implicitly assume that item attributes will explain all variability in discrimination parameters as we have not added an error term. Adding an error term for discrimination parameters using Bayesian estimation procedures would likely increase model fit and precision of standard error estimation. De Boeck (2008) has proposed a model with random error in item difficulty parameters for estimating models with crossed-random person and item effects with the package *lme*4 (Bates, Mächler, Bolker, & Walker, 2015) in *R*. Asparouhov and Muthen (2012) proposed an estimation procedure using a Bayesian methodology in Mplus for models with random effects for discrimination parameters in factor analysis models (see also Cho et al., 2014, for an explanatory approach with random residuals). Hence, future

research may further extend response style models using constrained discrimination parameters as proposed in explanatory IRT models that include random components of item or discrimination parameters.

Besides advancing estimation and modeling approaches, a substantive analysis of discrimination parameters of response style dimensions may help to identify sources of biases and problematic items in test construction. Discrimination parameters indicate item-specific differences in the strength of response style effects on item responses and hence indicate which items are more strongly affected by response style traits (see Falk & Cai, 2016). Testing hypotheses about moderating item attributes, such as ambiguity, item position, or complexity will provide valuable information to identify problematic items in test construction, or on finding an adequate test length for longer survey studies (see also Shao, Li, & Cheng, 2016, for a change point analysis to detect speeded responding). Such statistically and empirically informed item selection can lead to a reduction of the systematic impact of response styles on category choices and therewith biases in social science measurement situations (Podsakoff, MacKenzie, & Podsakoff, 2012).

Conclusion

The integration of Divide-by-Total IRT models that have accommodated response styles in different ways (Henninger & Meiser, 2019) highlighted commonalities, differences between, and implications of restrictions and specifications of the different IRT models. By making such differences and implications explicit, the suggested framework provides guidance for model selection in applied research.

In the applications of the framework in this article, latent covariances were crucial for model fit and items were impacted differently by response style dimensions in the Big Five standardization sample. Motivated by these results, we proposed two novel model extensions wherein the impact of response styles can vary, first, for different thresholds or categories within items, or, second, between items as a function of item attributes. The results from the empirical analysis and the development of two new

models illustrate how psychometric models can be used for test construction and to further develop theory on response styles.

Psychometric modeling of response styles is a useful tool to correct for and investigate biases in rating data. Furthermore, it allows us to test specific hypotheses through the comparison of alternative models. With the integration of various Divide-by-Total models in a common superordinate framework, we provide the basis to compare existing IRT models, choose the appropriate, or derive new variants in order to answer a wide variety of research questions.

References

Abad, F. J., Sorrel, M. A., Garcia, L. F., & Aluja, A. (2018). Modeling general, specific, and method variance in personality measures: Results for ZKA-PQ and NEO-PI-R. *Assessment*, *25*, 959–977. doi: 10.1177/1073191116667547

Asparouhov, T., & Muthen, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. In *Annual meeting of the national council on measurement in education* (pp. 1–27). Paper presented at the third UK Mplus Users' Meeting, London, UK. Retrieved from `http://www.statmodel.com/download/NCME12.pdf`

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi: 10.18637/jss.v067.i01

Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. doi: 10.1509/jmkr.38.2.143.18840

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(4), 608–628. doi: 10.1207/S15328007SEM0704_5

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

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

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

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical & Statistical Psychology*, *70*, 159–181. doi: 10.1111/bmsp.12086

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. doi: 10.1177/0146621608329891

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

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833. doi: 10.1177/0013164410388411

Borkenau, P., & Ostendorf, F. (2008). *NEO-Fünf-Faktoren Inventar nach Costa und McCrae (NEO-FFI).* Manual (2. Auflage). Göttingen: Hogrefe.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. doi: 10.1007/BF02294533

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. doi: 10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36–52. doi: 10.1037/a0030641

Cabooter, E. (2010). *The impact of situational and dispositional variables on response styles with respect to attitude measures.* Unpublished Doctoral Dissertation, Ghent, Belgium. Retrieved from `https://biblio.ugent.be/publication/4333765/file/4427719`

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

Cho, S. J., De Boeck, P., Embretson, S., & Rabe-Hesketh, S. (2014). Additive multilevel item structure models with random residuals: Item modeling for exaplanation and item generation. *Psychometrika*, *79*, 84–104. doi: 10.1007/s11336-013-9360-2

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. doi:

10.1007/s11336-008-9092-x

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*, 1–28. doi: 10.18637/jss.v048.c01

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Jong, M. G., Steenkamp, J.-B. E., 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*(1), 104–115. doi: 10.1509/jmkr.45.1.104

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20–30. doi: 10.1027//1015-5759.16.1.20

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433. doi: 10.1007/BF02294564

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

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

Hamilton, D. L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, *69*, 192–203. doi: 10.1037/h0025606

Henninger, M., & Meiser, T. (2019). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Accepted for Publication in Psychological Methods Pending on Minor Revisions*.

Huggins-Manley, A. C., & Algina, J. (2015). The Partial Credit Model and Generalized Partial Credit Model as constrained Nominal Response Models, with applications in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*, 308–318. doi: 10.1080/10705511.2014.937374

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response styles. *Journal of the American Statistical Association*, *102*, 454–463. doi: 10.1198/016214506000000960

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. doi: 10.3758/s13428-015-0631-y

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*, 116–138. doi: 10.1177/0013164413498876

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

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

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. doi: 10.1080/00273171.2013.866536

Kiefer, T., Robitzsch, A., & Wu, M. (2017). *TAM: Test analysis modules (Version 2.8-21) [Computer software].* Retrieved from `http://cran.r-project.org/package=TAM`

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance for dimensionality assessment for multidimensional models. *Journal of Educational Measurement*, *52*, 144–158. doi: 10.1111/jedm.12070

Lubbe, D., & Schuster, C. (2017). The graded response differential discrimination

model accounting for extreme response style. *Multivariate Behavioral Research*, 1–14. doi: 10.1080/00273171.2017.1350561

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

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362. doi: 10.1037/1082-989X.11.4.344

McKeown, B., & Thomas, D. (1988). *Q methodology.* Thousand Oaks, CA: Sage Publications.

Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment*, *24*, 27–34. doi: 10.1027/1015-5759.24.1.27

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical & Statistical Psychology.* doi: 10.1111/bmsp.12158

Moors, G. (2003). Diagnosing response style behavior by means of a latent-class factor approach. Socio-demographic correlates of gender role attitudes and perceptions of ethnic discrimination reexamined. *Quality and Quantity*, *37*, 277–302. doi: 10.1023/A:1024472110002

Morren, M., Gelissen, J. P., & 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. doi: 10.1111/j.1467-9531.2011.01238.x

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Pan, J., & Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics and Data Analysis*, *51*, 5765–5775. doi: 10.1016/j.csda.2006.10.003

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

10.1177/0013164416636655

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*, 633–654. doi: 10.1080/00273171.2018.1469966

Plieninger, H., Henninger, M., & Meiser, T. (2019). An experimental comparison of the effect of different response formats on response styles. *Manuscript submitted for publication*.

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. doi: 10.1177/0013164413514998

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual review of psychology*, *63*, 539–69. doi: 10.1146/annurev-psych-120710-100452

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.r-project.org`

Rollock, D., & Lui, P. P. (2016). Measurement invariance and the Five-Factor model of personality: Asian international and Euro American cultural groups. *Assessment*, *23*, 571–587. doi: 10.1177/1073191115590854

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. doi: 10.1198/016214501750332668

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*, 75–92. doi: 10.1111/j.2044-8317.1991.tb00951.x

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores [Monograph]. *Psychometrika*, *34*(Suppl. 17), 1–100. Retrieved from `https://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf`

Schmukle, S. C., Back, M. D., & Egloff, B. (2008). Validity of the five-factor model for the implicit self-concept of personality. *European Journal of Psychological Assessment*, *24*, 263–272. doi: 10.1027/1015-5759.24.4.263

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi: 10.1214/aos/1176344136

Schweizer, K. (2008). Investigating experimental effects within the framework of structural equation modeling: An example with effects on both error scores and reaction times. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 327–345. doi: 10.1080/10705510801922621

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141. doi: 10.1007/s11336-015-9476-7

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi: 10.1007/BF02294363

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567–577. doi: 10.1007/BF02295596

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. doi: 10.3102/1076998613481500

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. doi: 10.1086/214483

Tutz, G. (1997). Sequential models for ordered responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–142). New York: Springer.

Tutz, G., & Berger, M. (2016). Response styles in rating scales: Simultaneous modeling of content-related effects and the tendency to middle or extreme categories. *Journal of Educational and Behavioral Statistics*, *41*, 239–268. doi: 10.3102/1076998616636850

Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the Partial Credit

Model. *Applied Psychological Measurement*. doi: 10.1177/0146621617748322

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*, 195–217. doi: 10.1093/ijpor/eds021

Verhelst, N., Glas, C. A. W., & De Vries, H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–139). New York: Springer.

von Davier, M., Shin, H. J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*, *42*, 291–306. doi: 10.1177/0146621617730389

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. doi: 10.1111/j.1745-3984.2006.00020.x

Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, *48*, 441–456. doi: 10.1111/j.1745-3984.2011.00154.x

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. doi: 10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. doi: 10.1037/a0018721

Wetzel, E. (2013). *Investigation response styles and item homogeneity using Item Response Theory* (Doctoral dissertation). Retrieved from `http://d-nb.info/1058478389/34`

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

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*, 352–364. doi:

10.1027/1015-5759/a000291

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279–291. doi: 10.1177/1073191115583714

Zeller, F., Reiß, S., & Schweizer, K. (2017). Is the Item-Position Effect in Achievement Measures Induced by Increasing Item Difficulty? *Structural Equation Modeling: A Multidisciplinary Journal*, *24*, 745–754. doi: 10.1080/10705511.2017.1306706

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of Personality*, *84*, 461–472. doi: 10.1111/jopy.12172

Table 1

Exemplary Scoring Weights for an Item With 5 Response Categories

	Category number				
Primary Trait					
$\boldsymbol{s}^{\theta_n}$	0	1	2	3	4
$\boldsymbol{s}^{\theta_n}_{reversed-coded}$	4	3	2	1	0
Random Thresholds					
(*e.g., Wang et al., 2006*)					
$\boldsymbol{s}^{\theta^{\delta}_{n1}}$	0	1	1	1	1
$\boldsymbol{s}^{\theta^{\delta}_{n2}}$	0	0	1	1	1
$\boldsymbol{s}^{\theta^{\delta}_{n3}}$	0	0	0	1	1
$\boldsymbol{s}^{\theta^{\delta}_{n4}}$	0	0	0	0	1
Exploratory Response Styles					
(*e.g., Bolt & Johnson, 2009*)					
$\boldsymbol{s}^{\theta^{RS}_n}$	λ_0	λ_1	λ_2	λ_3	λ_4
Category Preferences (Sum-to-Zero)					
(*e.g., Bolt et al., 2014*)					
$\boldsymbol{s}^{\theta^{*}_{n1}}$	-1	1	0	0	0
$\boldsymbol{s}^{\theta^{*}_{n2}}$	-1	0	1	0	0
$\boldsymbol{s}^{\theta^{*}_{n3}}$	-1	0	0	1	0
$\boldsymbol{s}^{\theta^{*}_{n4}}$	-1	0	0	0	1
A Priori Specified Response Styles					
(*e.g., Wetzel & Carstensen, 2017*)					
$\boldsymbol{s}^{\theta^{ERS}_n}$	1	0	0	0	1
$\boldsymbol{s}^{\theta^{MRS}_n}$	0	0	1	0	0
$\boldsymbol{s}^{\theta^{ARS}_n}$	0	0	0	1	1
Proportional Effects of Response Styles					
(*New Variant*)					
$\boldsymbol{s}^{\theta^{ERS}_n}$	1	0	0	0	λ^{ERS}
$\boldsymbol{s}^{\theta^{MRS}_n}$	0	λ^{MRS}	1	λ^{MRS}	0
$\boldsymbol{s}^{\theta^{ARS}_n}$	0	0	0	1	λ^{ARS}

Note. ERS: Extreme Response Style, MRS: Mid Response Style, ARS: Acquiescence Response Style; EMRS: Extreme versus Mid Response Style; further scoring weight options: $EMRS_1 = (2, 1, 0, 1, 2)$, $EMRS_2 = (0, 1.5, 2, 1.5, 0)$, $ARS_2 = (0, 0, 0, 1, 2)$; adapted from Falk & Cai (2016); Tutz & Berger (2016); Weijters et al. (2010b); Wetzel & Carstensen (2017).

Table 2

Overview of Estimated Parameters and Model Fit Indices for the IRT Models

	Number of estimated						Model fit indices				
	parameters (total)	item-threshold parameters	latent variances	latent covariances	discrimination parameters		SGDDM	Log-Likelihood	BIC	LR-Test	Rank
Partial Credit Model	255	$60 \times 4 = 240$	5 (0)	10	0		.049	-880,496	1,763,381	—	8
Gen. Partial Credit Model	310	$60 \times 4 = 240$	5 (5) *(all fixed to 1)*	10	60	*60 Big Five*	.044	-873,093	1,749,091	—	7
Random Threshold Model *(Wang et al., 2006)*	259	$60 \times 4 = 240$	$5 + 4$ (0)	10	0		.047	-862,891	1,728,209	$\chi^2(4) = 35,210$ $p < .001$	6
Gen. Random Threshold Model *(based on Wang & Wu, 2011)*	370	$60 \times 4 = 240$	$5 + 4$ (9) *(all fixed to 1)*	10	120	*60 Big Five,* *60 Var. Thres.*	.041	-858,002	1,719,471	$\chi^2(60) = 30,182$ $p < .001$	5
Multidimensional NRM *(Bolt & Johnson, 2009)*	270	$60 \times 4 = 240$	$5 + 3$ (3) *(RS dim. fixed to 1)*	10	15	*5 × 3* *(5 for each RS)*	.046	-852,924	1,708,378	$\chi^2(15) = 55,143$ $p < .001$	3
Category Preference Model *(Bolt et al., 2014)*	285	$60 \times 4 = 240$	$5 + 4$ (0)	36	0		.046	-853,457	1,709,585	$\chi^2(30) = 54,077$ $p < .001$	4
Multidimensional PCM *(Bolt & Newton, 2011;* *Wetzel & Carstensen, 2017)*	276	$60 \times 4 = 240$	$5 + 3$ (0)	28	0		.046	-852,832	1,708,250	$\chi^2(21) = 55,327$ $p < .001$	2
Gen. Multidimensional PCM *(Falk & Cai, 2016)*	449	$60 \times 4 = 240$	$5 + 3$ (8) *(all fixed to 1)*	28	181	*60 Big Five,* *60 ERS/MRS,* *1 ARS*	.042	-845,521	1,695,248	$\chi^2(139) = 55,145$ $p < .001$	1

Note. Model estimation for Big Five personality factors with 60 items and $K + 1 = 5$ response categories; 5 Big Five plus response style dimensions if applicable; the number in parentheses indicates the number of latent variances fixed to 1; PCM: Partial Credit Model, NRM: Nominal Response Model, Gen.: Generalized; RS: Response Style, ERS: Extreme Response Style, MRS: Mid Response Style, ARS: Acquiescence Response Style, SGDDM: Standardized Generalized Dimensionality Discrepancy Measure, BIC: Bayesian Information Criterion, LR-Test: Likelihood Ratio Test (compare response style models to the PCM, gen. response style models to the gen. PCM), Rank based on BIC.

Table 3

Estimated Correlation Matrix in the Best Fitting Model (Generalized Multidimensional PCM by Falk & Cai, 2016; Variance of Latent Traits was Fixed to 1)

	Neuro.	Extra.	Open.	Agree.	Consc.	ERS	MRS	ARS
Neuroticism	1.00							
Extraversion	-0.45	1.00						
Openness	0.03	0.13	1.00					
Agreeableness	-0.13	0.26	0.05	1.00				
Conscientiousness	-0.32	0.13	-0.15	0.18	1.00			
ERS	0.11	-0.08	-0.08	-0.26	-0.10	1.00		
MRS	0.01	0.04	0.11	0.05	0.06	-0.15	1.00	
ARS	0.13	-0.04	-0.04	-0.28	-0.04	0.04	0.35	1.00

Note. Neuro: Neuroticism, Extra: Extraversion, Open: Openness, Agree: Agreeableness, Consc: Conscientiousness, ERS: Extreme Response Style, MRS: Mid Response Style, ARS: Acquiescence Response Style.

Table 4

Coding of Item Attributes for the Prediction of Discrimination Parameters α_{id} in the Explanatory IRT Modeling Extension

Item Coding			Intercept	Effects		
Negation	Complexity	Position	$\alpha_{Intercept}$	$\alpha_{Negation}$	$\alpha_{Complexity}$	$\alpha_{Position}$
—	—	—	1	0	0	0
✓	—	—	1	1	0	0
—	✓	—	1	0	1	0
✓	✓	—	1	1	1	0
—	—	✓	1	0	0	1
✓	—	✓	1	1	0	1
—	✓	✓	1	0	1	1
✓	✓	✓	1	1	1	1

Note. See also Table B1 in Appendix B for the coding for each of the 60 Big Five items.

Table 5

Overview of Estimated Parameters and Information Criteria for two new Exemplary Model Extensions

	Number of estimated					Model fit indices			
	parameters (total)	item-threshold parameters	latent variances	latent covariances	discrimination parameters	SGDDM	Log-Likelihood	BIC	LR-Test
Multidimensional PCM *(Lifting Equality Constraints from ARS Scoring Weights)*	277	$60 \times 4 = 240$	$5 + 3$ (0)	28	1 *1 ARS, (highest category)*	.046	-852,762	1,708,118	$\chi^2(1) = 141$ $p < .001$
Gen. Multidimensional PCM *(with Discrimination Parameters Explained by Item Attributes)*	282	$60 \times 4 = 240$	$5 + 3$ (2) *(ERS/MRS fixed to 1)*	28	8 *4 ERS, 4 MRS,*	.046	-852,467	1,707,576	$\chi^2(6) = 731$ $p < .001$

Note. Estimation of the two new model extensions for the Big Five personality factors with 60 items and $K + 1 = 5$ response categories; 5 Big Five plus response style dimensions; the number in parentheses indicates the number of latent variances that is fixed to 1; ERS: Extreme Response Style, MRS: Mid Response Style, ARS: Acquiescence Response Style; Style; SGDDM: Standardized Generalized Dimensionality Discrepancy Measure, BIC: Bayesian Information Criterion, LR-Test: Likelihood Ratio Test (compare the model to the multidimensional PCM by Bolt & Newton, 2011 or Wetzel & Carstensen, 2017, see Table 2).

Table 6

Estimated Elementary Parameters for α_{id} (with Standard Errors in Parentheses) in the Explanatory IRT Model with Response Style Dimensions

	Intercept	Negation	Complexity	Position
ERS	$1.02^{\dagger}$ (< 0.01)	$0.05^{\dagger}$ (0.01)	$-0.05^{\dagger}$ (0.01)	$0.13^{\dagger}$ (0.01)
MRS	$0.59^{\dagger}$ (< 0.01)	$0.06^{\dagger}$ (0.01)	-0.01 (0.01)	$0.06^{\dagger}$ (0.01)

Note. ERS: Extreme Response Style, MRS: Mid Response Style. Discrimination parameters are fixed to one for the primary trait dimensions; † $p < .001$.

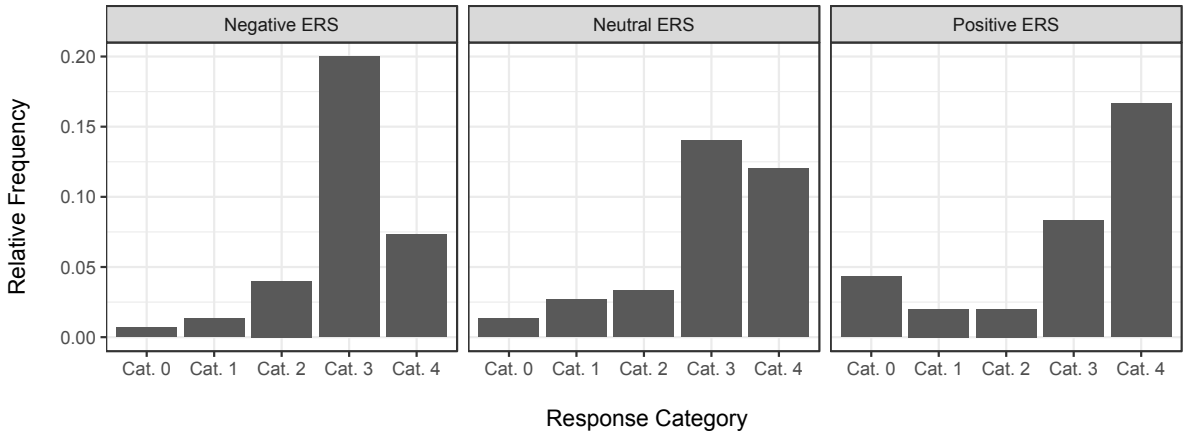

Figure 1. Relative frequencies of response category choices for three exemplary respondents based on simulated data with the same manifest mean across items ($\bar{X} = 3$, moderately positive trait levels), but different Extreme Response Style (ERS) levels.

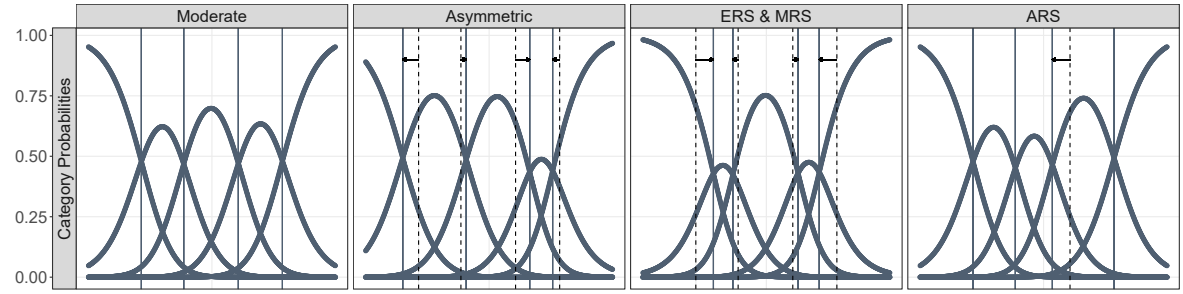

Figure 2. Illustration of category probability curves for an item i with five response categories $k \in \{0, ..., 4\}$. From left to right: for moderate respondents, respondents with a unique profile of asymmetric threshold shifts, respondents with positive Extreme and Mid Response Style (ERS & MRS), and respondents with positive Acquiescence Response Style (ARS).

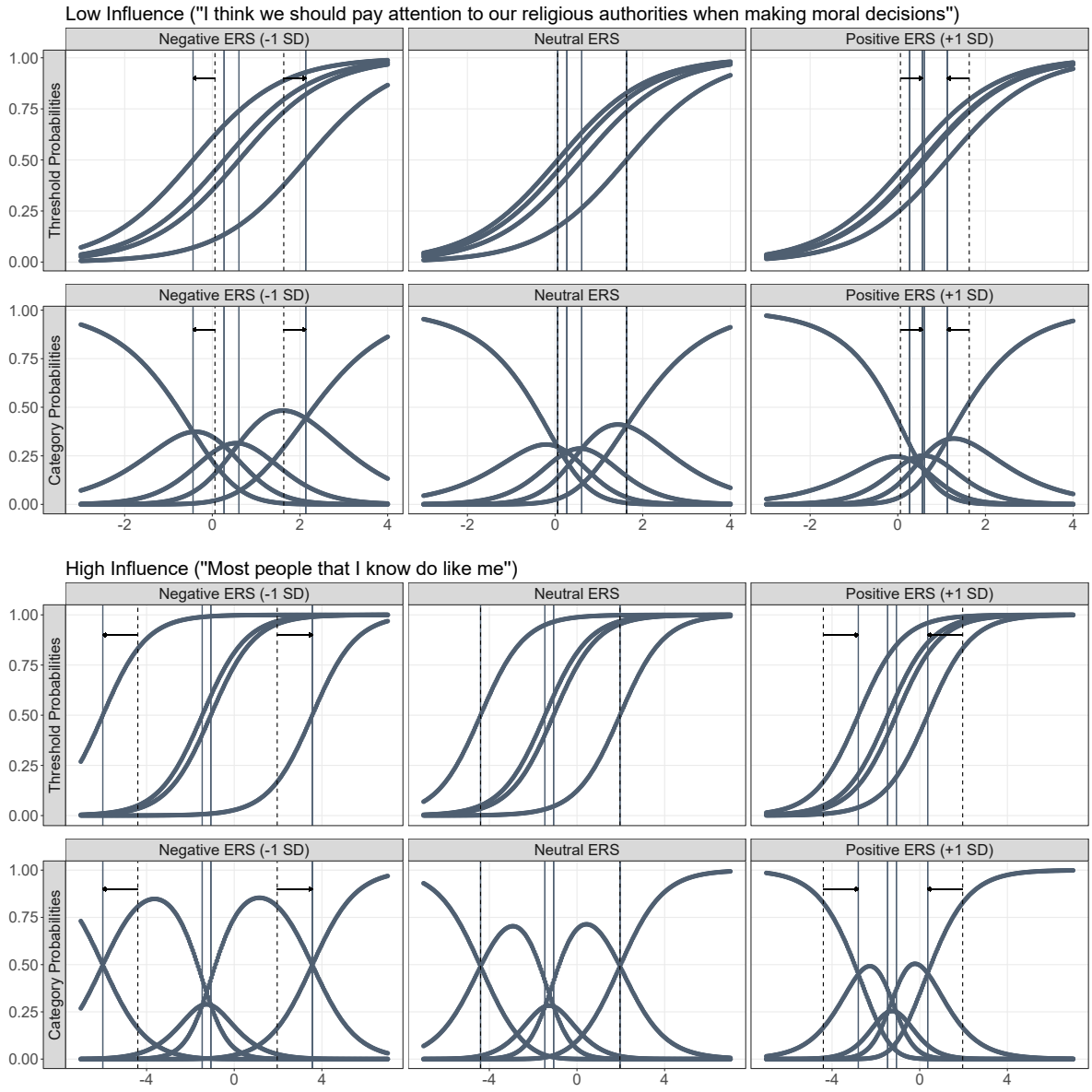

Figure 3. Illustration of the influence of low (upper panel) and high (lower panel) discriminability of the Extreme Response Style (ERS) dimension on threshold and category probabilities with model based item-threshold and discrimination parameters (Falk & Cai, 2016).

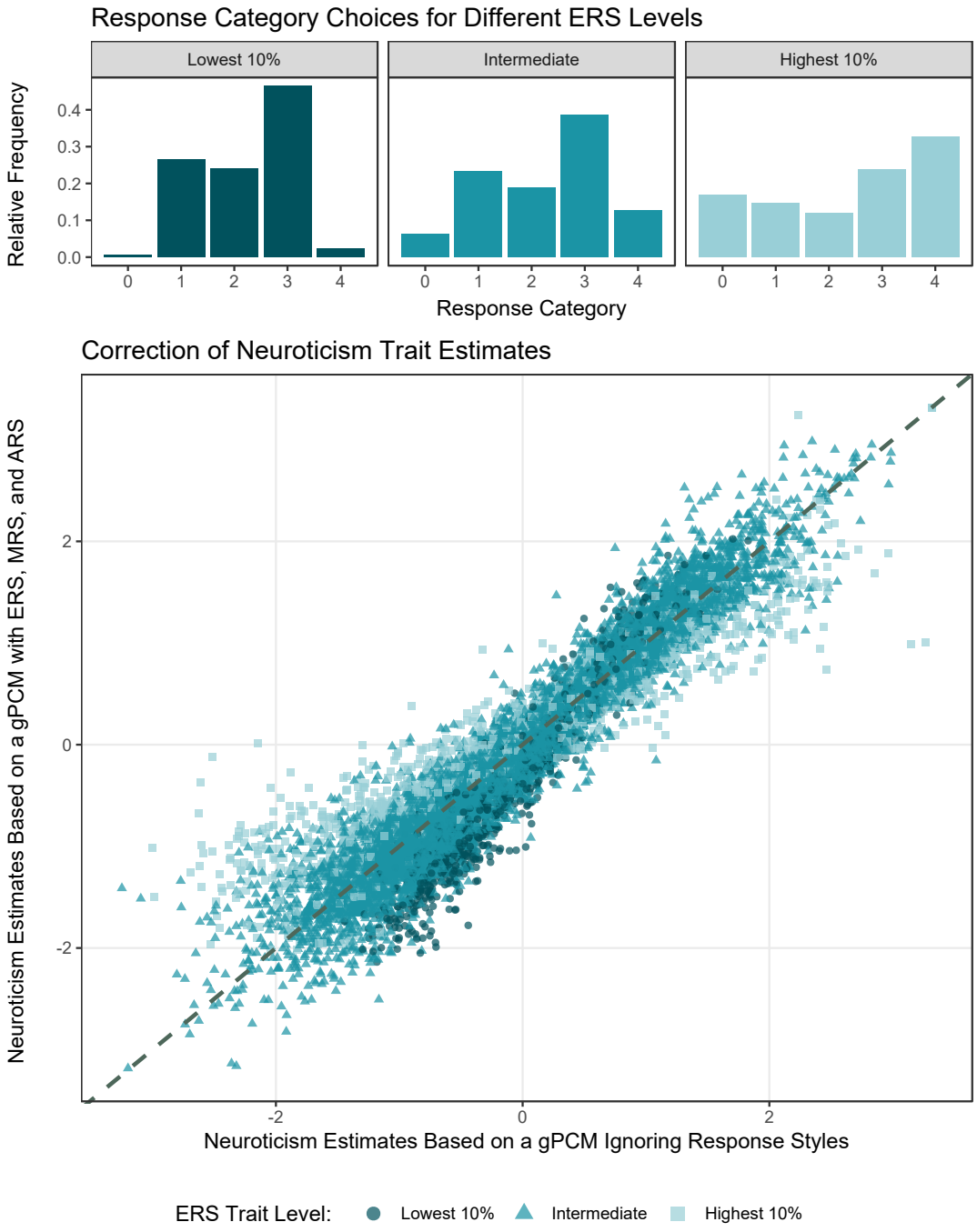

Figure 4. Upper panel: relative frequency of response category choices for lower and upper 10% quantiles and intermediate levels of Extreme Response Styles (ERS); lower panel: correction for Neuroticism estimates based on the generalized Partial Credit Model (gPCM) ignoring response styles and a gPCM with additional ERS, Mid Response Style (MRS), and Acquiescence Response Style (ARS) dimensions (Falk & Cai, 2016).

Appendix A

Exemplary Scoring Matrix for Two Primary Traits and Two Response Style Dimensions

Trait 1

	Cat 0	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	0	1	2	3	4
Item 2	0	1	2	3	4
Item 3	4	3	2	1	0
Item 4	4	3	2	1	0
Item 5	0	0	0	0	0
Item 6	0	0	0	0	0
Item 7	0	0	0	0	0
Item 8	0	0	0	0	0

Trait 2

	Cat 0	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	0	0	0	0	0
Item 2	0	0	0	0	0
Item 3	0	0	0	0	0
Item 4	0	0	0	0	0
Item 5	0	1	2	3	4
Item 6	0	1	2	3	4
Item 7	4	3	2	1	0
Item 8	4	3	2	1	0

ERS

	Cat 0	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	1	0	0	0	1
Item 2	1	0	0	0	1
Item 3	1	0	0	0	1
Item 4	1	0	0	0	1
Item 5	1	0	0	0	1
Item 6	1	0	0	0	1
Item 7	1	0	0	0	1
Item 8	1	0	0	0	1

MRS

	Cat 0	Cat 1	Cat 2	Cat 3	Cat 4
Item 1	0	0	1	0	0
Item 2	0	0	1	0	0
Item 3	0	0	1	0	0
Item 4	0	0	1	0	0
Item 5	0	0	1	0	0
Item 6	0	0	1	0	0
Item 7	0	0	1	0	0
Item 8	0	0	1	0	0

Appendix B

Coding of Item Characteristics for the 60 Big Five Items

Table B1

Coding Matrix for the Generalized Multidimensional PCM with Item Attribute Predictors for Discrimination Parameters of Response Style Dimensions

Item 1 - 30					Item 31 - 60				
Item	Neg.	Comp.	Pos.	Param.	Item	Neg.	Comp.	Pos.	Param.
N 1	1	0	0	$\alpha_I + \alpha_N$	N 31	0	0	0	α_I
E 2	0	0	0	α_I	E 32	0	0	0	α_I
O 3	1	0	0	$\alpha_I + \alpha_N$	O 33	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
A 4	0	1	0	$\alpha_I + \alpha_C$	A 34	0	0	1	$\alpha_I + \alpha_P$
C 5	0	0	0	α_I	C 35	0	0	0	α_I
N 6	0	0	1	$\alpha_I + \alpha_P$	N 36	0	1	0	$\alpha_I + \alpha_C$
E 7	0	0	1	$\alpha_I + \alpha_P$	E 37	0	0	1	$\alpha_I + \alpha_P$
O 8	0	0	0	α_I	O 38	0	1	0	$\alpha_I + \alpha_C$
A 9	0	0	1	$\alpha_I + \alpha_P$	A 39	0	0	0	α_I
C 10	0	1	0	$\alpha_I + \alpha_C$	C 40	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
N 11	0	1	0	$\alpha_I + \alpha_C$	N 41	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
E 12	1	0	1	$\alpha_I + \alpha_N + \alpha_P$	E 42	1	0	0	$\alpha_I + \alpha_N$
O 13	0	1	0	$\alpha_I + \alpha_C$	O 43	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
A 14	0	0	0	α_I	A 44	0	0	0	α_I
C 15	1	0	0	$\alpha_I + \alpha_N$	C 45	1	1	0	$\alpha_I + \alpha_N + \alpha_C$
N 16	0	0	0	α_I	N 46	0	0	0	α_I
E 17	0	0	1	$\alpha_I + \alpha_P$	E 47	0	0	1	$\alpha_I + \alpha_P$
O 18	0	1	0	$\alpha_I + \alpha_C$	O 48	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
A 19	0	1	0	$\alpha_I + \alpha_C$	A 49	0	0	0	α_I
C 20	0	0	0	α_I	C 50	0	1	0	$\alpha_I + \alpha_C$
N 21	0	0	0	α_I	N 51	0	1	0	$\alpha_I + \alpha_C$
E 22	0	0	1	$\alpha_I + \alpha_P$	E 52	0	0	1	$\alpha_I + \alpha_P$
O 23	1	0	1	$\alpha_I + \alpha_N + \alpha_P$	O 53	0	0	1	$\alpha_I + \alpha_P$
A 24	0	0	0	α_I	A 54	1	1	0	$\alpha_I + \alpha_N + \alpha_C$
C 25	0	1	0	$\alpha_I + \alpha_C$	C 55	1	1	1	$\alpha_I + \alpha_N + \alpha_C + \alpha_P$
N 26	0	0	0	α_I	N 56	0	1	0	$\alpha_I + \alpha_C$
E 27	0	0	0	α_I	E 57	0	1	1	$\alpha_I + \alpha_C + \alpha_P$
O 28	0	0	0	α_I	O 58	0	0	0	α_I
A 29	0	1	0	$\alpha_I + \alpha_C$	A 59	0	1	0	$\alpha_I + \alpha_C$
C 30	0	1	0	$\alpha_I + \alpha_C$	C 60	0	0	1	$\alpha_I + \alpha_P$

Note. Neg.: Negation; Comp.: Complexity, Pos.: Position (based on the 240 item measure), Param.: Parameter; N: Neuroticism; E: Extraversion; O: Openness; A: Agreeableness; C: Conscientiousness; α_I: α of the intercept, α_N: α for negated items, α_C: α for complex items, α_P: α for items appearing in the second half of the questionnaire.