Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration

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). 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

A large variety of Item Response Theory (IRT) modeling approaches aim at measuring and correcting for response styles in rating data. Here, we integrate response style models of the Divide-by-Total model family into one superordinate framework that parameterizes response styles as person-specific shifts in threshold parameters. This superordinate framework allows us to structure and compare existing approaches to modeling response styles and therewith makes model-implied restrictions explicit. With a simulation study, we show how the new framework allows us to assess consequences of violations of model assumptions and to compare response style estimates across different model parameterizations. The integrative framework of Divide-by-Total modeling approaches facilitates the correction for and examination of response styles. In addition to providing a superordinate framework for psychometric research, it gives guidance to applied researchers for model selection and specification in psychological assessment.

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

Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration

Many researchers use rating scales to assess latent variables such as beliefs, attitudes or personality traits. Rating scales are in widespread use as they are convenient to apply and evaluate. However, rating responses do not only capture the primary trait (i.e. the trait to be measured), but also other sources of interindividual differences. Respondents might use satisficing strategies when retrieving knowledge from memory (Krosnick, 1991), rely on contextual cues (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), answer in a socially desirable way (Ellingson, Smith, & Sacket, 2001), or show preferences for certain response categories (e.g., Paulhus, 1991; Van Vaerenbergh & Thomas, 2013). If respondents use the rating scale in different manners, these differences are inherent in their responses to rating scale items besides the trait that is intended to be measured. In consequence, inferences for psychological assessment or research questions that are drawn from rating data are prone to be biased when interindividual differences in response tendencies are ignored.

One such source of interindividual differences in rating scale usage are response styles, respondents' tendencies to prefer specific kinds of categories over others. For example, a tendency towards choosing the highest and lowest categories is called *extreme response style* (ERS), a tendency towards the middle category is called *mid response style* (MRS), and a tendency to generally agree or disagree with an item is called *acquiescence* (ARS) or *disacquiescence* (DARS), respectively (for a review see Van Vaerenbergh & Thomas, 2013). Research found response styles to be consistent across traits (Weijters, Geuens, & Schillewaert, 2010a; Wetzel, Carstensen, & Böhnke, 2013), and stable over time (Weijters, Geuens, & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2016).

Although Plieninger (2017) showed in a simulation study that under certain conditions response styles had only minor effects on traditional measures of test quality such as Cronbach's alpha, ignoring response styles can distort inferences drawn from measurement: for example, a respondent with a tendency for extreme categories may

receive a higher or lower trait estimate than a respondent with a moderate preference for extreme categories (e.g., Bolt, Lu, & Kim, 2014; Meiser & Machunsky, 2008). Ignoring response styles can also distort relationships between measured variables. To give an example, Böckenholt and Meiser (2017) illustrated that the relation between latent dimensions was inflated when response styles were ignored. Accounting for response styles is also relevant when comparing different subgroups, such as age, gender or cultural backgrounds. For example, it has been shown in the context of cross-cultural research that respondents from different countries vary in their use of the rating scale. This differential usage of the rating scale biases inferences on cultural differences when response tendencies are not accounted for (e.g., Bolt et al., 2014; Cheung & Rensvold, 2000; Morren, Gelissen, & Vermunt, 2012).

Many psychometric modeling approaches have been proposed in order to measure and control for response styles in ordinal rating data. Response styles have been accommodated in various types of Item Response Theory (IRT) models such as extensions of Divide-by-Total models (e.g., Bolt & Johnson, 2009; Falk & Cai, 2016; Rost, 1991; Wang, Wilson, & Shih, 2006; Wetzel & Carstensen, 2017), the Graded Response Model (GRM, e.g., Ferrando, 2014; Lubbe & Schuster, 2017; Rossi, Gilula, & Allenby, 2001; Thissen-Roe & Thissen, 2013), and IRTree models that characterize responses to a rating scale item by a sequence of a priori defined multiple processes (Böckenholt, 2012; De Boeck & Partchev, 2012; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014). The psychometric models differ in the degree of a priori assumptions on response styles that they incorporate. While some are constructed to account for predefined response styles such as ERS or MRS (e.g., Böckenholt, 2012; De Jong, Steenkamp, Fox, & Baumgartner, 2008; Falk & Cai, 2016; Jin & Wang, 2014; Johnson, 2003; Lubbe & Schuster, 2017; Morren, Gelissen, & Vermunt, 2011; Rossi et al., 2001; Thissen-Roe & Thissen, 2013; Wetzel & Carstensen, 2017), others aim to correct for heterogeneity in response scale use without a priori assumptions on the nature of response styles (e.g., Bolt & Johnson, 2009; Moors, 2003; Rost, 1991; Wang et al., 2006). Besides, the models also differ in whether they formalize response styles as

discrete parameters that give rise to subpopulations (as is the case in latent class analyses, e.g., Moors, 2003; Morren et al., 2011; Rost, 1991), or as continuous parameters that are reflected by additional traits (e.g., Böckenholt, 2012; Bolt & Johnson, 2009; Wang et al., 2006; Wetzel & Carstensen, 2017). They also differ with regard to whether they conceptualize response styles as additional person parameters (e.g., Böckenholt, 2012; Bolt & Johnson, 2009; Moors, 2003; Wetzel & Carstensen, 2017) or heterogeneity in item-specific threshold parameters (e.g., Jin & Wang, 2014; Rost, 1991; Wang et al., 2006).

This article focuses on psychometric model variants for response styles in the framework of Divide-by-Total IRT models for ordinal rating data. In Divide-by-Total models, the exponential of the parameter combination of one category is divided by the sum across all categories (see Thissen & Steinberg, 1986). In Divide-by-Total models for ordinal responses, a threshold parameter indicates the value on the latent continuum for which two adjacent response categories are equally likely, such that the category probability curves intersect. In consequence, response style effects can be illustrated as shifts in the thresholds that have a direct effect on threshold locations and category probabilities of the two neighboring categories and the response distribution as a whole. In contrast, GRMs characterize category probabilities as the integral under the density function between two adjacent thresholds so that the location and difference between two thresholds render the probability of choosing a certain category. Therefore, thresholds have a different meaning in GRMs than in Divide-by-Total models and do not fall into the class of models considered here. As a third model class, IRTree models decompose rating responses into multiple subprocesses corresponding to a priori defined judgment processes. In contrast to IRTrees, Divide-by-Total models allow for exploratory as well as confirmatory analyses of response styles. Furthermore, IRTree models often dichotomize indicators of the latent trait (see De Boeck & Partchev, 2012, for more details and alternative types of IRTree models). In consequence, the intensity of category choice (e.g., choosing "strongly agree" instead of "agree") in an IRTree model may solely be determined through response styles and does not involve the primary

trait to be measured (although there are IRTree extensions allowing for simultaneous effects of the primary trait and response styles in one node of the IRTree model, see Jeon & De Boeck, 2016; Meiser, Plieninger, & Henninger, 2019). Divide-by-Total models, however, retain the ordinal response process for the trait and can model response styles as additional trait dimensions or as shifts of thresholds. For these reasons, and because most of the prominent models for response styles fall into this class, extensions of Divide-by-Total models for response styles, rather than GRMs or IRTree models, are the focus of the present article.

Our goal is to integrate the different modeling approaches into one superordinate framework that combines two lines of literature that have extended Divide-by-Total models to incorporate response styles either in terms of variations in thresholds or in terms of additional trait dimensions. For this purpose, we present one common formalization of response style parameters, structure the models based on assumptions that they make on response styles, and show commonalities and differences between the response style models. In a simulation study, we show the benefit of using a joint framework for response style effects to compare estimates of response styles across modeling approaches. The superordinate framework will support researchers in examining the theoretical and empirical differences between existing response style models and in assessing the added value when developing new model variants.

A Superordinate Framework of IRT Models for Response Styles

The models considered in this article are IRT-based modeling approaches for response styles and their factor analytic equivalent of the family of Divide-by-Total models (Thissen & Steinberg, 1986): the *Nominal Response Model* (NRM, Bock, 1972; Takane & de Leeuw, 1987), special cases for ordinal items such as the *Partial Credit Model* (PCM, Masters, 1982), and *Rating Scale Model* (RSM, Andrich, 1978) as well as the *Generalized Partial Credit Model* with item-specific discrimination parameters (gPCM, e.g., Muraki, 1992, see also Mellenbergh, 1995).

In Divide-by-Total IRT models, response styles can be illustrated by the location

of threshold parameters and category probability curves. The left column of Figure 1 shows the threshold characteristic curves (upper row) and category probability curves (lower row) for one exemplary item with five response categories $k \in \{0, ..., 4\}$ and four equally spaced thresholds under an ordinal Divide-by-Total model for respondents with moderate response styles. The threshold probability curves display the conditional probability of choosing category k given that the response is either in category $k-1$ or k, while the category probability curves display the probability that person n chooses category k of item i as a function of the latent person parameter. The vertical lines in both graphs depict the $K = 4$ thresholds. In ordinal Divide-by-Total models with ordered thresholds, the category probabilities of two adjacent categories $k-1$ and k are equal at threshold k, where the threshold probability equals .5 and the category probability curves intersect (see Figure 1).

INSERT FIGURE 1 ABOUT HERE

The threshold probability is given by

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

and is as a function of the trait parameter θ_n for person n and the item-specific category parameter b_{ik} for item i and category k.

The category probability formula of a Divide-by-Total model for $K+1$ categories with $k \in \{0, ..., K\}$ (a PCM adapted from Masters, 1982) is given by

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

In Divide-by-Total models, category probabilities are set as ratios of the exponential of a linear parameter combination divided by its sum across all categories ensuring that the category probabilities sum to 1. Consequently, the single category probabilities are interdependent such that the probability for one category depends on the parameters of all other categories. The category or scoring weights $\boldsymbol{s}_k$ describe the

relation between trait and category. They can be estimated in the NRM (by using a sum-to-zero constraint within items or by setting the weight of one category to 0), as opposed to being fixed, for example to $\boldsymbol{s} = (0, ..., K)$, in the PCM. The item-specific category parameter b_{ik} can be decomposed into an item location β_i and thresholds τ_{ik}, with $b_{ik} = \beta_i + \tau_{ik}$ and $\beta_i = (\sum_{k=1}^{K} b_{ik})/K$. When threshold parameters are equal for all items ($\tau_{ik} = \tau_k$), the model reduces to a RSM. For identification, the parameters of the first category in Equation 2 are set to 0 ($s_0\theta_n - b_{i0} \equiv 0$). In generalized models, item-specific discrimination parameters α_i indicate the impact of the latent dimension θ_n on the item response through the linear parameter combination $\alpha_i s_k \theta_n - \sum_{k'=0}^{k} b_{ik'}$ (Muraki, 1992).

The Divide-by-Total models in Equation 1 and 2 do not incorporate response style effects. The main assumption underlying such IRT models is that covariation between item responses is solely due to the underlying trait. This requirement is the basis for drawing inferences on respondents' latent traits from scale scores. However, when response styles are present, they influence item responses besides the latent trait and introduce additional covariance between items. In consequence, additional person or item parameters must be added to account for this covariance.

Modeling Response Styles as Varying Thresholds or Additional Traits

To account for response style variance in rating scale data, different extensions of Divide-by-Total models have been presented in the literature. They differ in how they specify response styles, namely as variation in thresholds or additional person traits. The two perspectives exist side-by-side, however they represent two lines of literature that are rarely connected to each other (but see Rijmen & De Boeck, 2005, for a similar approach).

Taking a threshold-based perspective, response styles can be seen as variation in the thresholds that capture remaining covariation between items conditional on the trait (e.g., Jin & Wang, 2014; Rost, 1991; Wang et al., 2006; Wang & Wu, 2011). This perspective is based on the reasoning that the assumption of homogeneous threshold

parameters is violated, so that thresholds must be allowed to vary between respondents or subpopulations of respondents. For example, ERS manifests itself by shifting the upper and lower thresholds towards the item location, increasing the probability of choosing the highest and lowest category (see column 2 in Figure 1).

From a trait-based perspective, one can extend the IRT model to a multidimensional model and include an additional trait parameter for each response style (ERS, MRS, ARS, or specific category preferences, e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017). These additional traits reflect that respondents differ in their tendencies to prefer specific kinds of categories over others and thus use the rating scale heterogeneously. For example, a person with positive ERS trait levels has a tendency to choose extreme over intermediate categories, and vice versa for low ERS trait levels (see column 2 in Figure 1).

Formalizing Response Styles as Person-Specific Threshold Shifts

Our goal is to connect the two lines of literature and to integrate the different psychometric models for response styles into one common, superordinate framework. In this framework, response styles can be equivalently seen as varying thresholds or as additional traits and are parameterized as person-specific shifts in the thresholds. Consider the threshold (upper row) and category (lower row) probability curves of an ordinal Divide-by-Total model in Figure 1. Both, threshold and category probability curves reflect response styles through shifts in the thresholds. When ERS is positive, the outer thresholds move inwards, when MRS is positive, the inner thresholds move outwards and vice versa for negative ERS or MRS, respectively. When ARS is positive, the threshold separating the middle category and the first agreement category is shifted to the left, increasing the probability that the response is given in one of the two agreement categories. Independent of whether the model defines response styles as variations in thresholds or additional trait parameters, both perspectives on response styles can be reconciled in parameterizing response styles as person-specific shifts in threshold parameters. Therefore, we propose a superordinate modeling framework in

which we define threshold and category probabilities as

$$p(X = k|X \in \{k-1, k\}, \theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(\theta_n - b_{ik} + \delta_{nk}\right)}{1 + \exp\left(\theta_n - b_{ik} + \delta_{nk}\right)} \tag{3}$$

and

$$p(X = k|\theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k} b_{ik'} + \sum_{k'=0}^{k} \delta_{nk'}\right)}{\sum_{j=0}^{K} \exp\left(s_j\theta_n - \sum_{k'=0}^{j} b_{ik'} + \sum_{k'=0}^{j} \delta_{nk'}\right)} \tag{4}$$

with $s_0\theta_n - b_{i0} + \delta_{n0} \equiv 0$. Herein, θ_n is the respondent's trait parameter and δ_{nk} a parameter of a person-specific shift in threshold k with $[\theta, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. As before, b_{ik} is the item-specific category parameter for item i and category k with $b_{ik} = \beta_i + \tau_{ik}$ for $k \in \{0, ..., K\}$ and scoring weights $\boldsymbol{s}_k$ reflect the relation between trait and category[1]. Even though person-specific threshold shift parameters δ_{nk} influence the location of one specific threshold k separating two adjacent categories $k-1$ and k, all category probabilities are impacted as the denominator of the model is defined as the sum across categories (see Equation 4).

Please note that δ_{nk} can be seen as a *person-specific shift of threshold parameter k*, but also as a *threshold-specific person parameter*: seeing δ_{nk} as a person-specific shift of threshold parameter k, quantifying the interindividual deviance from the item threshold due to response tendencies towards either category k or $k-1$, we can rewrite the linear parameter combination in Equation 3 as $\theta_n + (\delta_{nk} - b_{ik})$. Considering δ_{nk} to be a threshold-specific person parameter that for a specific threshold adds to or subtracts from the trait parameter of the respondent and therewith reflects his or her tendency to prefer certain categories over others, we can rewrite the linear parameter combination

[1] Under certain conditions, person-specific threshold shifts may also be item-specific (δ_{nik}, e.g., Jin & Wang, 2014), and some modeling approaches propose generalizations of this framework using discrimination parameters for primary trait θ_n and person-specific threshold shifts δ_{nk} (Falk & Cai, 2016; Wang & Wu, 2011). Here, we refrained from adding the index i (δ_{nik}) and discrimination parameters (α_{id}) to the general framework in order to avoid additional complexity (but see Table 1 and Tables A1 and A2 in Appendix A).

as $(\theta_n + \delta_{nk}) - b_{ik}$. Thus, we can take a threshold-based or person-based perspective on response styles within one IRT model formulation (c.f. Rijmen & De Boeck, 2005, for a comparison between multidimensional IRT models and mixture models through shift parameters).

Of course, the modeling framework in Equations 3 and 4 is not identified as primary trait θ_n and person-specific thresholds δ_{nk} cannot be separated. The modeling approaches in the literature have identified special cases from this superordinate framework by either putting restrictions on response styles δ_{nk}, covariance matrix $\mathbf{\Sigma}$, or both. To define a special case from the superordinate framework, one must initially specify how response styles are expected to shift the thresholds, that is the composition of person-specific thresholds δ_{nk}. For example, in case that one aims at modeling ERS, threshold shifts of the outer thresholds are expected to be symmetric around the item location (see Figure 1). Then, one must evaluate whether person-specific threshold shifts δ_{nk} are still redundant to the latent primary trait(s): they are not redundant when, for example, ERS is modeled, however, they are redundant when all thresholds potentially shift into one direction. To achieve separability of primary trait(s) θ_n and person-specific threshold shifts δ_{nk}, one must either put (further) restrictions on response style effects δ_{nk} or constrain the variance-covariance matrix $\mathbf{\Sigma}$.

To facilitate model estimation, response styles can additionally be modeled through extraneous item sets (i.e. items other than those measuring the primary traits, e.g., Wetzel & Carstensen, 2017) or anchoring vignettes (short passages describing hypothetical scenarios that respondents must rate using rating scales; for examples see Bolt et al., 2014). Similarly, models for response styles including a linear pattern (e.g., ARS whose coding goes along with the trait) or little a priori assumptions (e.g., Bolt & Johnson, 2009; Bolt et al., 2014) require the inclusion of reversed coded items to reliably separate trait and response styles. Another option is constraining response styles to be equal for several scales, hence modeling general response tendencies across different content domains (e.g., Bolt & Newton, 2011; Moors, 2003; Weijters et al., 2010a; Wetzel & Carstensen, 2017).

Model Integration

We now demonstrate how different variants of response style IRT models from the Divide-by-Total model family in the literature have specified response styles (i.e., person-specific threshold shifts) δ_{nk}, hence which restrictions were put on δ_{nk} and/or $\mathbf{\Sigma}$. For each modeling approach, we show the linear parameter combination used to model primary trait θ_n, item-threshold parameter b_{ik} and response styles δ_{nk}.

When response styles are specified as variations in the thresholds, commonly a threshold probability notation (or logit notation) was applied by the respective authors (see Equation 3; e.g., Jin & Wang, 2014; Wang et al., 2006; Wang & Wu, 2011). In contrast, when response styles are specified as additional traits, a category probability formulation (usually including category scoring weights) was commonly used (see Equation 4; e.g., Bolt & Johnson, 2009; Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017). Of course, we can reformulate threshold probabilities in terms of category probabilities and vice versa. In the former case, we cumulate the linear predictor across categories. With such a reformulation from threshold to category probabilities, we can derive cumulative scoring weights for latent trait and response style dimensions (e.g., $\boldsymbol{s}^{trait} = (0, 1, 2, 3, 4)$, $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$, see the following section on model equivalence using the notation of multidimensional NRMs and Appendix B). In the latter case, threshold probabilities can be computed from category probabilities according to

$$p(X = k|X \in \{k-1, k\}, \theta, \boldsymbol{b}) = \frac{P(X = k)}{P(X = k-1)} \Big/ \left(1 + \frac{P(X = k)}{P(X = k-1)}\right). \tag{5}$$

In practice, this amounts to reversing the cumulation by subtracting the parameters of category $k-1$ from the parameters of category k to obtain the linear predictor of the threshold probability notation (as an example, see the decumulation in the simulation study further below). Converting category probabilities into threshold probabilities is a helpful tool in Divide-by-Total models to examine the effects that response styles have on specific thresholds[2].

[2] Please note that we use s for cumulative scoring weights in the category probability notation (see

Independent of whether the IRT models accounting for response styles specify response styles as varying thresholds or additional traits, we structure the modeling approaches proposed in the literature in three groups. In the first group, the respective models assume that person-specific thresholds are independent from each other and from the latent trait. In the second group, the models constrain person-specific threshold shifts so that response style effects are captured by latent classes or additional response style dimensions. To separate trait from response style effects, the variance-covariance matrix of trait and response style dimensions is typically constrained to a diagonal matrix. In the third group of models, response styles are defined a priori, for example through fixing scoring weights of response style dimensions. This allows one to estimate the full variance-covariance matrix between primary trait and response style dimensions. In Table 1, we give an overview of the three groups of models and highlight whether they take a threshold- or trait-based perspective on response styles, the assumed distribution of response style parameters and response style specification, exemplary research questions that can be answered with the respective model, the linear predictor of the model and further model characteristics. For more details on the notation of model formulas, see Tables A1 and A2 in Appendix A. In addition, we illustrate instances of threshold shifts in each group of models for four exemplary respondents in Figure 2.

INSERT TABLE 1 ABOUT HERE

Models Assuming Independent Person-Specific Threshold Shifts

The first group of modeling approaches accounts for unknown response styles in the data. Each respondent has a unique individual threshold-shift profile (see upper row in Figure 2 and section 1 in Table 1), as person-specific threshold shifts are considered independent from each other.

Wang and colleagues (Wang et al., 2006; Wang & Wu, 2011) proposed such a

---

Equation 4), and s^* for scoring weights adapted to the threshold probability notation (not cumulated across categories; see e.g., Table 1 and Table A2 in Appendix A).

varying threshold approach using the linear predictor $\theta_n - (\beta_i + \tau_{ik} - \delta_{nk})$. Hence, each respondent is characterized by his or her own threshold shift parameters δ_{nk} that increase probabilities for certain, while decreasing probabilities for other categories. Large variances of the threshold shift parameters across respondents indicate pronounced variability between persons in their response tendencies; in case that all threshold variances equal 0, the model reduces to a PCM or RSM. In order to disentangle the primary trait from person-specific shifts in the thresholds and to identify this specific response style model from the general framework (Equation 3 and 4), Wang and colleagues restricted the variance-covariance matrix $\mathbf{\Sigma}$ of trait and varying thresholds to a diagonal matrix and thus assumed uncorrelated trait and threshold effects. The assumption of independent threshold shifts, however, is violated when response styles such as ERS or MRS that require symmetric threshold shifts around the item location (see columns 2 and 3 in Figure 1) are present in the data. Wang and Wu (2011) extended the IRT model to incorporate item-specific discrimination parameters $\alpha_i(\theta_n - (\beta_i + \tau_{ik} - \delta_{nk}))$ describing the relation between items and random effects for persons $[\theta, \delta_1, ..., \delta_K]$.

INSERT FIGURE 2 ABOUT HERE

Models Constraining Person-Specific Threshold Shifts, but Estimating Response Styles Exploratorily

In the second group of models, response styles are not specified a priori, but systematics between threshold shifts across persons can be modeled. The middle row in Figure 2 illustrates category probability curves for four exemplary respondents in a multidimensional NRM with estimated scoring weights for one response style dimension. Hence, these models search for a common structure of threshold shifts across respondents in the data: we see that the profile of threshold shifts is equal across respondents, while the magnitude and direction differs between respondents. Models belonging to this group are mixture distribution models (Böckenholt & Meiser, 2017; Moors, 2003; Rost, 1991) and multidimensional NRMs (Bolt & Johnson, 2009; Bolt et

al., 2014, see section 2 in Table 1)[3].

Mixture distribution models. Rost (1991) proposed an extension of the PCM to a latent class or mixture distribution model (for applications see Austin, Deary, & Egan, 2006; Eid & Rauber, 2000; Gollwitzer, Eid, & Jürgensen, 2005; Meiser & Machunsky, 2008; Wetzel, Carstensen, & Böhnke, 2013, see also von Davier & Rost, 2006). Mixtures of PCMs account for heterogeneity in response scale use by identifying latent subpopulations. The polytomous Rasch model is assumed to hold within each subpopulation c with subpopulation specific item and threshold parameters $\theta_{cn} - b_{cik}$ accounting for different response tendencies between the subpopulations. Hence, response styles are assumed to be homogeneous within, but heterogeneous between latent subpopulations. Many applications of the mixture distribution model have consistently suggested the existence of two subpopulations: one subpopulation with moderate response style and another subpopulation with ERS in which thresholds are shifted towards the item location (e.g., Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel, Carstensen, & Böhnke, 2013). In order to disentangle parameters $\beta_i + \tau_{ik}$ that are constant across subpopulations and threshold shifts δ_{ck} that quantify the subpopulation-specific shift in threshold k, one can decompose $b_{cik} = \beta_i + \tau_{ik} + \delta_{ck}$ (see Meiser & Machunsky, 2008; Wetzel, Böhnke, Carstensen, Ziegler, & Ostendorf, 2013). Theoretically, as the number of classes approaches the number of respondents, this model is equivalent to a model with person-specific threshold shifts (see Equation 3). In its latent class form, it restricts response styles to be discrete latent variables.

Latent class models account for response styles in an exploratory manner and at the cost of additional parameters to be estimated. In order to introduce more parsimonious and confirmatory model variants, Böckenholt and Meiser (2017) proposed a linear function describing distances between adjacent thresholds across latent subpopulations. For instance, threshold distances for respondents in subpopulation 2

---

[3] Please note that although we illustrate threshold shifts for one response style dimension in Figure 2, it is also possible to model multiple independent response style dimensions in the multidimensional NRM leading to more individualized threshold shift profiles.

can be defined as a linear function of threshold distances in subpopulation 1. Then, $\delta_{1k} = \delta_{1(k-1)} = 0$ holds for subpopulation 1, while the threshold distances in subpopulation 2 are specified as $(\tau_{ik} + \delta_{2k}) - (\tau_{i(k-1)} + \delta_{2(k-1)}) = a + b(\tau_{1ik} - \tau_{1i(k-1)})$.

The trait-based counterpart to latent class mixture models for response styles was proposed by Moors (2003). Similar to Rost (1991), Moors modeled one additional response style with discrete levels using latent class factor analysis with a logit link. Here, the item-specific category parameter b_{ik} is represented by the intercept in the factor model, while scoring weights and traits $s_{dk}\theta_{nd}$ are represented by slopes and factors, respectively for each of the D dimensions. Hence, the linear predictor in the model by Moors is given by $\sum_{d=1}^{D} \theta_{nd} - b_{ik} + s_k^{*RS}\theta_n^{RS}$, wherein the superscript ^{RS} flags the response style trait. Moors (2003) used fixed ordinal scoring weights for primary traits and estimated category scoring weights for one response style dimension freely. As scoring weights were positive for the extreme categories, but negative for the intermediate categories, it seems that ERS is present in the data.

Multidimensional Nominal Response Models. Bolt and Johnson (2009) extended the NRM (Bock, 1972; Takane & de Leeuw, 1987) to a multidimensional model for a trait and D response styles RS with $\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$. In contrast to latent class analyses, they conceptualized response styles as continuous traits in the IRT model. The category scoring weights s_{dk}^{RS} for response styles can be estimated and interpreted post hoc: Similar to the interpretation by Moors (2003), positive scoring weights for the two extreme categories and negative weights for the intermediate categories indicate ERS. When scoring weights are estimated, the covariance matrix of the multivariate trait distribution (trait and response style dimensions) is restricted to an identity matrix for identification, implying that latent dimensions are uncorrelated (Bolt & Johnson, 2009, see also Johnson & Bolt, 2010).

A general model for response tendencies based on the multidimensional NRM was proposed by Bolt et al. (2014). They modeled response styles as person-specific preferences θ_{nk}^{RS} for each of the $K + 1$ categories using the linear predictor $\theta_n - b_{ik} + \theta_{nk}^{*RS}$. The category-specific response style traits θ_{nk}^{RS} describe the tendency of

respondents to choose category k across items. Bolt and colleagues fixed the scoring weights for primary traits and estimated person-specific preferences for categories. The model for category-specific response tendencies θ_{nk}^{RS} can be reformulated into a model using person-specific threshold shifts δ_{nk}. Then person-specific threshold shifts are composed of the category preferences of the two adjacent categories bounding the respective threshold: $\delta_{nk} = \theta_{nk}^{*RS} = \theta_{nk}^{RS} - \theta_{n(k-1)}^{RS}$. Bolt et al. (2014) used a sum-to-zero constraint for the response style traits across categories within persons and anchoring vignettes to separate response styles from traits. The variance-covariance matrix of random effects was estimated and correlations between category preference parameters guide the interpretation of response style effects. For example, correlations of the category preference parameters of the extreme categories suggest an ERS effect.

Models Using A Priori Specifications of Response Styles

The models in the last group use a priori specifications of response styles. These specifications entail restrictions on threshold shifts, and fix the structure of threshold shifts a priori. The lower row in Figure 2 illustrates threshold shifts for a multidimensional PCM with two response style dimensions (ERS, affecting Thresholds 1 and 4 and MRS, affecting Thresholds 2 and 3). We can see that threshold shifts are symmetric around the item location, and that each respondent has a unique combination of the impact of ERS and MRS on threshold shifts (e.g., Respondent 1 has large ERS, but essentially no MRS shifts, while Respondent 2 has small negative ERS and MRS shifts). A threshold dispersion model (Jin & Wang, 2014), a constrained variant of a mixture distribution model (Morren et al., 2011), and multidimensional extensions of the PCM (Bolt & Newton, 2011; Wetzel & Carstensen, 2017) or generalized PCM (Falk & Cai, 2016) belong to this group of models (see section 3 in Table 1).

Jin and Wang (2014) modified the random threshold model by Wang and colleagues to account for ERS. Instead of modeling K person-specific threshold parameters, they introduced one person-specific weight parameter θ_n^W for all thresholds

with a lognormal distribution using the linear predictor $\theta_n - (\beta_i + \theta_n^w \tau_{ik})$. The parameter θ_n^W can be interpreted as a person-specific threshold dispersion parameter: it pulls apart the thresholds when $\theta_n^W > 1$, decreasing the probability for extreme categories, and pushes the thresholds together when $\theta_n^W < 1$, increasing the probability for extreme categories. In order to reparameterize Jin and Wang's approach in terms of person-specific shifts in threshold parameters, we can disentangle the term $\theta_n - (\beta_i + \theta_n^W \tau_{ik})$ into $\theta_n - (\beta_i + \tau_{ik}) - \tau_{ik}(\theta_n^W - 1)$. This separates thresholds τ_{ik} that are equal for all respondents and respondent-specific threshold shifts $\delta_{nik} = -\tau_{ik}(\theta_n^W - 1)$ varying between respondents.

Morren et al. (2011) extended the approach by Moors (2003) and showed that restrictions of the scoring weights for response styles allow for the inclusion of theoretical assumptions, such as a tendency for extreme categories (through $\boldsymbol{s}_k^{ERS} = (1.5, -1, -1, -1, 1.5)$. Hence, the latent class factor models can also be seen as a constrained variant of the multidimensional NRM by Bolt and colleagues ($\theta_n - b_{ik} + s_k^{*RS}\theta_n^{RS}$; Bolt & Johnson, 2009; Bolt & Newton, 2011) with a priori specified scoring weights for the response style trait. The models differ insofar as Moors (2003) and Morren et al. (2011) assumed that the latent response style trait is a variable with discrete levels, while Bolt and colleagues conceptualize response styles as continuous traits.

Multidimensional (Generalized) Partial Credit Models. Bolt and Newton (2011) as well as Wetzel and Carstensen (2017) used the multidimensional NRM and PCM (Rasch, 1961, see also Kelderman, 1996; Meiser, 1996) to model the primary trait and theoretically defined response styles such as ERS, MRS, and ARS ($\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$). For that purpose, they fixed category scoring weights for the trait and response styles (e.g., $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$, $\boldsymbol{s}^{MRS} = (0, 0, 1, 0, 0)$, $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ for an item with 5 response categories). For example, through $\boldsymbol{s}^{ERS}$ the ERS trait describes how much the outer thresholds move inwards for positive θ_n^{ERS} and outwards for negative θ_n^{ERS}. As the scoring weights are equal for the lowest and highest category, the threshold pair (1 and 4) is perfectly negatively correlated:

$\theta_n^{ERS} = -\delta_{n1} = \delta_{n4}$ (see also column 2 in Figure 1 and Appendix B). Tutz, Schauberger, and Berger (2018) proposed another special case of a multidimensional PCM wherein a response style trait is weighted by a scaling factor that is a function of the number of response categories $\delta_{nk} = ({}^{K}/_{2} - k + 0.5)\ \theta_n^{RS}$ (for an odd number of categories)[4]. Hence, positive θ_n^{RS} imply a tendency towards the middle category and negative θ_n^{RS} a tendency towards extreme categories. Because scoring weights for different traits are fixed, the full variance-covariance matrix of trait and response style dimensions can be estimated. This allows researchers to investigate relations between primary traits and response styles.

Falk and Cai (2016) built on the work of Bolt and colleagues: they extended the multidimensional NRM to include discrimination parameters α_{id} indicating the relation between items i and latent dimension d across categories in the IRT model $(\alpha_i\theta_n - b_{ik} + \sum_{d=1}^{D}(\alpha_{id}^{RS} s_{dk}^{*RS})\theta_{nd}^{RS})$. Discrimination parameters α_{id} describe the relation between items and primary trait or response style dimensions. In the model by Falk and Cai (2016), person-specific threshold shifts δ_{nik} are composed of discrimination parameters α_{id}^{RS}, scoring weights s_{dk}^{RS}, and trait parameters θ_{nd}^{RS}. The authors also summarize different possibilities to estimate, constrain or fix scoring weights in a multidimensional NRM and provide an overview of constraints used with different IRT model variants. Through disentangling discrimination parameters (reflecting the relationship between the item and trait) from scoring weights (reflecting the relation between categories and traits), item-specific response style effects can be tested (for more details see Falk & Cai, 2016, p.332ff).

Model Equivalence in the Notation of $\boldsymbol{T}$ Matrices

The different model specifications in combination with identification constraints result in the large variety of different approaches to modeling response styles in the response style literature. We can subsume all models presented under Equations 3 and

[4] For example, person-specific thresholds shifts for a five category item are defined as $\boldsymbol{\delta}_n = (1.5 \cdot \theta_n^{RS}, 0.5 \cdot \theta_n^{RS}, -0.5 \cdot \theta_n^{RS}, -1.5 \cdot \theta_n^{RS})$, with cumulative scoring weights $\boldsymbol{s}^{RS} = (0, 1.5, 2, 1.5, 0)$.

4, as we can reformulate their varying threshold or additional trait specifications of response styles as person-specific threshold shifts with restrictions on δ_{nk} or $\boldsymbol{\Sigma}$.

Therefore, we can consider the superordinate framework for the various Divide-by-Total models in Equation 4 as a multidimensional extension of a NRM (Bock, 1972; Takane & de Leeuw, 1987). A framework to specify NRMs using a matrix notation was proposed by Thissen and Steinberg (1986). Here, we use this notational approach to describe how person-specific threshold shifts δ_{nk} are specified and restricted in the different models. This allows us to derive cumulative scoring weights for response style effects for all models that, in turn, are essential for model estimation in standard software such as Mplus (Muthén & Muthén, 2012) or in the statistical programming environment *R* (R Core Team, 2018) with packages *TAM* (Kiefer, Robitzsch, & Wu, 2017) or *mirt* (Chalmers, 2012) that use a multidimensional NRM parameterization of IRT models (see Henninger & Meiser, 2019, for a discussion on software implementation).

Thissen and Steinberg (1986) defined the category probability for person n and item i in a standard NRM—the cumulation of the linear predictor $\theta_n + b_{ik}$ across categories (see Equation 2)—through the k^{th} entry of $\boldsymbol{\alpha}' \times \boldsymbol{T}^a \theta_n + \boldsymbol{\gamma}_i' \times \boldsymbol{T}^c$, where $\boldsymbol{\alpha}'$ and $\boldsymbol{\gamma}'$ are parameter vectors of length K, while $\boldsymbol{T}^a$ and $\boldsymbol{T}^c$ represent two $K \times (K+1)$ design matrices (see Thissen & Steinberg, 1986, p. 571). We extend the linear parameter combination by δ_{nk} and thus add $\boldsymbol{\delta}_n \times \boldsymbol{T}^d$. Herein, $\boldsymbol{\delta}_n$ is the n^{th} row of a matrix of dimension $N \times K$ containing the person-specific threshold shift parameters of N persons and K thresholds. $\boldsymbol{T}^d$ is a $K \times (K+1)$ design matrix (see below). The design matrix $\boldsymbol{T}^d$ allows us to derive the cumulative scoring weights for certain types of person-specific threshold shifts, as specified in the different modeling approaches presented in the previous section.

In the superordinate framework that we propose (see Equation 3 and 4), the n^{th} row of the matrix $\boldsymbol{\delta}$ is given by

$$\boldsymbol{\delta}_n = \Big(\delta_{n1}, \quad \delta_{n2}, \quad \ldots, \quad \delta_{nK}\Big)$$

and $\boldsymbol{T}^d$ is a design matrix with dimensions $K \times (K+1)$ that cumulates person-specific

threshold shifts across categories:

$$\boldsymbol{T}^d = \begin{pmatrix} 0 & 1 & 1 & \dots & 1 \\ 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Hence, the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ is given by $\left(0, \quad \delta_{n1}, \quad \delta_{n1} + \delta_{n2}, \quad \dots, \quad \sum_{k=1}^{K} \delta_{nk}\right)$ which is equivalent to the cumulative sum of person-specific threshold shifts across categories for person n in the category probability notation (Equation 4). It follows that the design matrix $\boldsymbol{T}^d$ is a representation of the scoring weights for K person-specific threshold shift dimensions in the category probability notation.

Model with Person- or Subpopulation-Specific Threshold Shifts

In a random threshold model using varying thresholds for response style effects (RTM, e.g., Wang et al., 2006), $\boldsymbol{\delta}$ is a $N \times K$ matrix. To identify the model and separate trait from response style effects, the variance-covariance matrix $\boldsymbol{\Sigma}$ is constrained to a diagonal matrix. To reflect a mixture distribution model (Rost, 1991), the matrix $\boldsymbol{\delta}$ can be reduced to a matrix of dimensions $C \times K$, where C is the total number of latent classes. Hence, $\boldsymbol{\delta} \times \boldsymbol{T}^d$ results in a $C \times (K + 1)$ matrix, where the c^{th} row is given by $\left(0, \quad \delta_{c1}, \quad \delta_{c1} + \delta_{n2}, \quad \dots, \quad \sum_{k=1}^{K} \delta_{ck}\right)$.

Models Constraining Person-Specific Threshold Shifts

In order to elucidate the restrictions that multidimensional extensions of the NRM (Bolt & Johnson, 2009; Moors, 2003) impose on person-specific threshold shifts δ_{nk}, we illustrate the integration procedure for one additional response style dimension θ_n^{RS}. In this case, K person-specific threshold shifts δ_{nk} are condensed into one response style dimension θ_n^{RS}. In consequence, θ_n^{RS} is person-specific with regards to the magnitude of response style effects. Thresholds are differently affected through the inclusion of freely estimated scoring weights s_k that differ between categories, but are equal between

persons. As outlined in the model review, δ_{nk} is restricted to be a function of scoring weights s_k^* and the response style trait θ_n^{RS}. Therefore, the n^{th} row of the matrix $\boldsymbol{\delta}$ containing the person-specific threshold shifts for n persons and k thresholds is given by $\boldsymbol{\delta}_n = \Big(s_1^*\theta_n^{RS},\ \ s_2^*\theta_n^{RS},\ \ \ldots,\ \ s_K^*\theta_n^{RS}\Big)$.

In consequence, the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ is given by

$$\Big(0,\ \ s_1^*\theta_n^{RS},\ \ (s_1^* + s_2^*)\theta_n^{RS},\ \ \ldots,\ \ (\textstyle\sum_{k=1}^{K} s_k^*)\theta_n^{RS}\Big)$$

and the cumulative category scoring weights for the response style trait θ_n^{RS} are given by $\boldsymbol{s} = \Big(s_1^*,\ \ s_1^* + s_2^*,\ \ \ldots,\ \ \sum_{k=1}^{K} s_K^*\Big) = \Big(0,\ \ s_1,\ \ s_2,\ \ \ldots,\ \ s_K\Big)$. In case that θ_n^{RS} is discrete, we obtain the model by Moors (2003), whereas for continuous θ_n^{RS}, we obtain the model by Bolt and Johnson (2009).

Category Preference Model. A modeling approach wherein response styles are parameterized as $K + 1$ category preferences was proposed by Bolt et al. (2014). In this model, category preferences are not cumulated across thresholds, but solely affect the specific category. Therefore, we have to reverse the cumulative nature of category probabilities (see Equation 5) by defining $\delta_{nk} = \theta_{nk}^{RS} - \theta_{n(k-1)}^{RS}$. Hence, the n^{th} row of the $\boldsymbol{\delta}$ matrix is given by $\boldsymbol{\delta}_n = \Big(\theta_{n1}^{RS} - \theta_{n0}^{RS},\ \ \theta_{n2}^{RS} - \theta_{n1}^{RS},\ \ \ldots,\ \ \theta_{nK}^{RS} - \theta_{n(K-1)}^{RS}\Big)$ with $\theta_{n0}^{RS} \equiv 0$.

In consequence the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ is given by

$$\Big(0,\ \ \theta_{n1}^{RS},\ \ \theta_{n2}^{RS},\ \ \ldots,\ \ \theta_{nK}^{RS}\Big)$$

so that each category preference of each person (θ_{nk}^{RS}) is solely part of the linear parameter combination of category k (see also Bolt et al., 2014, or Table A2 in Appendix A).

Instead of restricting $\delta_{nk} = \theta_{nk}^{RS} - \theta_{n(k-1)}^{RS}$, we can also alter the design matrix in order to directly estimate the category preference parameter θ_{nk}^{RS}, a matrix wherein the

n^{th} row is given by $(\theta_{n1}^{RS}, ..., \theta_{nK}^{RS})$. For this purpose, the design matrix $\boldsymbol{T}^{d}$ is modified to

$$\boldsymbol{T}^{d*} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

so that the n^{th} row of the matrix $\boldsymbol{\theta} \times \boldsymbol{T}^{d*}$ is, in consequence, given by

$$\left(0, \quad \theta_{n1}^{RS}, \quad \theta_{n2}^{RS}, \quad ..., \quad \theta_{nK}^{RS}\right).$$

The model with response styles as category preferences separated the primary trait θ_n from category preferences θ_{nk}^{RS} by restricting the category preferences to sum to zero within respondents across categories. In order to include this restriction $\sum_{k=1}^{K} \theta_{nk} = 0$, we again alter the design matrix $\boldsymbol{T}^{d*}$ to the format

$$\boldsymbol{T}^{d**} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ -1 & 0 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ -1 & 0 & 0 & \dots & 1 \end{pmatrix}$$

so that the n^{th} row of $\boldsymbol{\theta} \times \boldsymbol{T}^{d**}$ is given by

$$\left(-\sum_{k=1}^{K} \theta_{nk}^{RS}, \quad \theta_{n1}^{RS}, \quad \theta_{n2}^{RS}, \quad \dots, \quad \theta_{nK}^{RS}\right)$$

and category preferences θ_{nk}^{RS} sum to zero within respondents across categories.

Models Using a Priori Specifications of Response Styles

Threshold Dispersion Model. Jin and Wang (2014) used a person-specific dispersion parameter θ_n^W that pulls thresholds τ_{ik} apart or pushes them together in order to account for ERS. Therefore, person-specific threshold shifts δ_{nik} are defined as a function of θ_n^W and τ_{ik} that can be disentangled into thresholds τ_{ik} that are fixed and

person-specific threshold shifts $\delta_{nik} = -\tau_{ik}(\theta_n^W - 1)$. For item i, the n^{th} row of the matrix $\boldsymbol{\delta}$ is given by $\boldsymbol{\delta}_{ni} = \Big(-\tau_{i1}(\theta_n^W - 1), \quad -\tau_{i2}(\theta_n^W - 1), \quad \ldots, \quad -\tau_{iK}(\theta_n^W - 1)\Big)$, and in consequence, the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ is given by

$$\Big(0, \quad -\tau_{i1}(\theta_n^W - 1), \quad -(\tau_{i1} + \tau_{i2}) \cdot (\theta_n^W - 1), \quad \ldots, \quad -(\textstyle\sum_{k=1}^{K} \tau_{ik}) \cdot (\theta_n^W - 1)\Big).$$

Multidimensional NRM / PCM. A multidimensional PCM for response styles (e.g., Bolt & Newton, 2011; Falk & Cai, 2016; Tutz et al., 2018; Wetzel & Carstensen, 2017) can be specified as a special case of the superordinate framework through imposing restrictions on δ_{nk}. Here, we demonstrate the restrictions on δ_{nk} for a model with three response style dimensions $\theta_n^{ERS}, \theta_n^{MRS}$, and θ_n^{ARS} and five response categories ($k \in \{0, ..., 4\}$). The scoring weights of the response style dimensions ($\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$, $\boldsymbol{s}^{MRS} = (0, 0, 1, 0, 0)$, and $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$) define which category is affected by which response style. The scoring weights are cumulative as they originate from the category probability formulation (Equation 4), but can be converted into adapted scoring weights s^* for threshold probabilities. As we have seen above, these adapted scoring weights are the difference between the scoring weights of two adjacent categories, so $\boldsymbol{s}^{*ERS} = (-1, 0, 0, 1)$, $\boldsymbol{s}^{*MRS} = (0, 1, -1, 0)$, and $\boldsymbol{s}^{*ARS} = (0, 0, 1, 0)$ as can also be seen in the threshold shifts in Figure 1 and Appendix B). Building upon scoring weights s_k^*, we see which thresholds are impacted by which response style trait. For example, the first threshold is impacted by $-\theta_n^{ERS}$, the second by θ_n^{MRS}, the third threshold by $-\theta_n^{MRS} + \theta_n^{ARS}$, while the fourth threshold is impacted by θ_n^{ERS}. Including these restrictions on δ_{nk}, the n^{th} row of the matrix $\boldsymbol{\delta}$ containing the response style effects on thresholds is given by $\boldsymbol{\delta}_n = \Big(-\theta_n^{ERS}, \quad \theta_n^{MRS}, \quad -\theta_n^{MRS} + \theta_n^{ARS}, \quad \theta_n^{ERS}\Big)$.

In consequence, the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ is given by

$$\Big(0, \quad -\theta_n^{ERS}, \quad -\theta_n^{ERS} + \theta_n^{MRS}, \quad -\theta_n^{ERS} + \theta_n^{ARS}, \quad \theta_n^{ARS}\Big).$$

From the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^d$ we can in turn see the scoring weights for response styles ERS, MRS, and ARS in a multidimensional PCM, as $\boldsymbol{s}^{ERS} = (0, -1, -1, -1, 0)$ or alternatively $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ are the scoring weights for the ERS latent trait, $\boldsymbol{s}^{MRS} = (0, 0, 1, 0, 0)$ for the MRS latent trait, and $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ for the ARS

latent trait that were specified this way in the original modeling approaches (e.g., Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017, see also Appendix B).

In conclusion, the different Divide-by-Total modeling extensions for response styles can be summarized in one common framework in which response styles are parameterized as person-specific threshold shifts. Thus, all the modeling approaches can be written in terms of threshold and category probabilities and regarded as extensions of the multidimensional NRM.

Simulation Study

We present a short simulation study to illustrate the benefits of integrating the different IRT models for response styles into one framework. As response style specifications differ between the modeling approaches, it has, on the one hand, not been obvious what kind of assumptions, specification, and restrictions were implemented in the models, and, on the other hand, how to compare estimates of response styles between IRT approaches. Our framework highlighted how response styles can be specified as person-specific threshold shifts δ_{nk} and which restrictions were implemented in the different response style models (e.g., the constraint on the covariance matrix by Wang et al., 2016, or the assumption of symmetry of threshold shifts for ERS by Wetzel & Carstensen, 2017). This allows us to analyze the sensitivity to violations of inherent assumptions in response style IRT models, and the goodness of parameter recovery with respect to primary trait and response style dimensions.

In the simulation study, we examined primary trait and response style parameter estimation of a selection of response style IRT models in scenarios with one ERS dimension that equally affects Thresholds 1 and 4 ($\boldsymbol{\delta}_n = (-\theta_n^{ERS}, 0, 0, \theta_n^{ERS})$), and different levels of covariation between threshold shifts and primary traits. The simulation study therefore allows us to (1) examine effects of varying covariation on parameter recovery and (2) illustrate response style parameter recovery in terms of person-specific threshold shifts.

Setup for Data Generation and Model Fit

We set the number of thresholds to $K = 4$, the number of respondents to $N = 500$, and the number of items to $I = 50$ with 25 items for each of two primary dimensions. In order to facilitate estimation of the response style models, each primary dimension contained 10 reversed-coded items. In each replication, item parameters were drawn from a truncated normal distribution $TN(0, 1, -1.5, 1.5)$ and centered, while threshold parameters were drawn from a uniform distribution $U(-2.5, 2.5)$, centered and ordered in ascending sequence. The variance of the two primary and one ERS dimension was fixed to 1, the covariance between the primary traits was fixed to $\rho = .2$, and for each replication the correlation between the primary traits and the ERS trait was drawn from a Wishart distribution with 5 degrees of freedom and set equal for the two primary dimensions. Respondents' trait parameters were generated from a $MVN \sim (\mathbf{0}, \mathbf{\Sigma})$.

In order to illustrate how to convert estimated response style parameters of the ERS dimension into person-specific threshold shifts of Threshold 1 and 4, we selected the following models: a PCM, a random threshold model (Wang et al., 2006), a multidimensional NRM (Bolt & Johnson, 2009), a model with person-specific category preferences (Bolt et al., 2014), and a multidimensional PCM (Wetzel & Carstensen, 2017). The random threshold model by Wang et al. (2006) already provides us with estimates of person-specific threshold shifts, but constrains these to be independent from each other and the primary traits. For the multidimensional NRM by Bolt and Johnson (2009) and PCM by Wetzel and Carstensen (2017), we used estimated or fixed scoring weights, respectively, to weigh the response style trait and subtracted the parameters for neighboring categories to obtain person-specific threshold shifts, $\delta_{nk} = s_k^* \theta_n^{RS} = (s_k - s_{(k-1)})\theta_n^{RS}$. Both models can account for the symmetric threshold shifts of the ERS dimension. But when scoring weights are estimated as in the multidimensional NRM (Bolt & Johnson, 2009), correlations between response styles and primary traits cannot be taken into account. Such correlations can be accounted for when scoring weights are fixed and $\mathbf{\Sigma}$ is estimated as in the multidimensional PCM (Wetzel & Carstensen, 2017). In the model by Bolt et al. (2014) we subtracted category

preferences of neighboring categories to obtain person-specific threshold shifts $\delta_{nk} = \theta^{*RS}_{nk} = \theta^{RS}_{nk} - \theta^{RS}_{n(k-1)}$. This model can account for the symmetric threshold shifts through ERS and correlations between primary traits and ERS. All model were estimated using R (R Core Team, 2018) with the package *TAM* (Test Analysis Modules, Kiefer et al., 2017) using marginal maximum likelihood method with a quasi Monte-Carlo integration procedure.

We realized $R = 5000$ replications and evaluated the estimation of trait and person-specific threshold shifts (Threshold 1 and Threshold 4) in terms of the correlation between true and estimated parameters ($\text{Cor} = r(\hat{\theta}_n, \theta_n)$) and mean bias ($\text{Bias} = \sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)/_N$) for each replication r.

Results and Conclusion

Figure 3 shows the correlation between true and estimated parameters and mean bias for the two primary traits (upper panel) and Threshold 1 and 4 (lower panel). In terms of correlation between true and estimated parameters, we see that response style models have a higher correlation of true and estimated primary trait parameters than the PCM that does not account for response styles. Overall, differences between models are small, and the minimum correlation between true and estimated primary trait parameters still amounts to $r = .95$ for the PCM. The correlation between true and estimated response style parameters is lower than for trait parameters. For primary trait and response style parameters, the random threshold model (Wang et al., 2006) has the lowest correlation within the response style models. This is not surprising given that it assumes independent latent dimensions ($\Sigma = Diag$) and was misspecified in this simulation scenario where $\rho(\delta_1, \delta_4) = -1$. Furthermore, for primary and response style traits, we see negative quadratic trends for models restricting the covariance between primary traits and threshold shifts to 0 (Bolt & Johnson, 2009; Wang et al., 2006), and positive quadratic trends for models estimating these correlations (Bolt et al., 2014; Wetzel & Carstensen, 2017).

INSERT FIGURE 3 ABOUT HERE

Even though bias is considerably small for all models, some systematic biases in terms of person parameter estimation can be seen in Figure 3. In the PCM, primary traits were overestimated for negative, and underestimated for positive correlations between primary traits and person-specific threshold shifts. For primary traits as well as person-specific threshold shifts, on average bias levels were smallest for the multidimensional NRM and multidimensional PCM (Bolt & Johnson, 2009; Wetzel & Carstensen, 2017), but worse for the random threshold model (Wang et al., 2006).

Overall, it seems that the multidimensional NRM (Bolt & Johnson, 2009) was relatively robust when primary traits and person-specific thresholds showed correlations in the population model, even if the model assumes independent latent dimensions. Unsurprisingly, the multidimensional PCM (Wetzel & Carstensen, 2017)—the data generating model—performed well in estimating primary trait and response style parameters. The simulation study illustrates how assumptions of response style models can be tested, and how estimates of response style parameters can be compared across models that have originally used different parameterizations. This simulation may inspire further analyses. For instance, one could evaluate the impact of skewed response distributions due to relatively easy or difficult items on the validity of primary trait or response style trait estimation. From a model application perspective, it would be helpful to examine the structure of data that is necessary to apply different response style models. For instance, how many different dimensions are necessary to obtain adequate heterogeneity within items to estimate primary and response style traits? Which models require reversed-coded items, for which models do they facilitate estimation? Further simulations of this kind will be the basis for evidence-based and rational model choices, in particular when not only traits, but response styles themselves become an object of study.

Discussion

We proposed a superordinate framework for various Divide-by-Total IRT models accounting for response styles. In this framework, response styles are modeled through

person-specific thresholds shift parameters. These parameters reflect differences in respondents' tendencies to prefer types of categories over others. We have demonstrated that numerous IRT modeling approaches for response styles proposed in the literature can be subsumed under this umbrella framework by restricting either person-specific threshold shifts δ_{nk}, the variance-covariance matrix of person effects $\mathbf{\Sigma}$, or both. This includes approaches modeling response styles as random noise (Wang et al., 2006; Wang & Wu, 2011), investigating response styles exploratorily (Böckenholt & Meiser, 2017; Bolt & Johnson, 2009; Moors, 2003; Rost, 1991), or defining response styles a priori (Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Jin & Wang, 2014; Morren et al., 2011; Wetzel & Carstensen, 2017, see Table 1 and Figure 2). Therewith, two lines of literature that have parameterized response styles either as variations in the thresholds (e.g., Jin & Wang, 2014; Rost, 1991; Wang et al., 2006), or as additional traits (e.g. Bolt & Johnson, 2009; Bolt et al., 2014; Falk & Cai, 2016; Moors, 2003; Wetzel & Carstensen, 2017) are integrated into one common framework.

Using the matrix notation by Thissen and Steinberg (1986), we showed that the different model variants can be considered as multidimensional extensions of a NRM using person-specific variations in thresholds to incorporate response styles. This integrative perspective on the numerous response style models with their different parameterizations highlights the restrictions on δ_{nk} and allows us to derive cumulative scoring weights for model estimation in a joint software framework such as Mplus (Muthén & Muthén, 2012, see also Huggins-Manley & Algina, 2015) or in the statistical programming environment *R* (R Core Team, 2018) with packages *TAM* (Kiefer et al., 2017) or *mirt* (Chalmers, 2012).

Furthermore, the integration of Divide-by-Total model extensions allows us to interpret response styles across different response style specifications. Translating response style traits into person-specific threshold shifts makes it possible to see how the various models capture response behavior. With the simulation study, we illustrated how effects of ERS (a shift in Thresholds 1 and 4) can be investigated across IRT model variants that, for example, have originally specified response styles as functions of

scoring weights and additional person dimensions.

Highlighting Model-Implied Effects of Response Style Parameterizations

The joint framework proposed here highlights the commonalities and differences between the existing modeling approaches and therewith illuminates the specific implications of each modeling approach. By translating scoring weights into person-specific shifts in the thresholds (and vice versa, see model review, section on matrix notation, the simulation study and Appendix B), the model-implied effects of response styles on threshold and category probabilities become visible. This reparameterization is particularly relevant for multidimensional models as it highlights how the scoring weights of response style traits translate into threshold shifts and which implications are implicitly made on threshold and category probabilities.

As an example, we see that a specification of ERS using scoring weights $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ implies symmetric person-specific threshold shifts of the first and last threshold around the item location $-\delta_{n1} = \delta_{n4}$, while the two intermediate thresholds are not affected by ERS. ARS is typically defined through scoring weights $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ which translates into a person-specific shift of the third threshold, while all other thresholds stay constant. Alternatively, ARS can be defined as a person-specific shift in thresholds 3 and 4 through $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 2)$ (see Henninger & Meiser, 2019, for a discussion of related models that map new theoretical assumptions on scoring weights). Adding response style parameters into the IRT model changes the distance between thresholds as these are shifted by δ_{nk} (see Figure 1 and Figure 2). However, a shift in the threshold affects the category probability of all categories, as in Divide-by-Total models the denominator of category probabilities is defined by the sum across all categories. So, even when some thresholds are not shifted, in the face of response styles, the probabilities of *all* categories change as a characteristic of Divide-by-Total models.

Implications and Outlook

Even though we restrict ourselves to response style models belonging to the Divide-by-Total model family (hence excluding models from the GRM, sequential, or IRTree model families), our unified framework integrates a variety of response style models with many different assumptions and characteristics (see Table 1). Divide-by-Total models are flexible tools as they allow for within-item multidimensionality of item responses and for the possibility to model response styles in an exploratory as well as confirmatory way. These possibilities result in a large variety of models. Being aware of these modeling options and their model-implied assumptions allows us to test specific restrictions on response styles while staying within the Divide-by-Total framework. Examining response style models within one IRT model family like Divide-by-Total models facilitates model comparisons for testing specific theoretical assumptions without confounds with the overall model structure.

Having integrated the various IRT model extensions for response styles into one unifying framework, the restrictions and assumptions that are imposed on response styles in each model become more explicit. Besides correcting for biases in rating data, psychometric modeling of response styles is a useful tool to test theoretical assumptions on response styles in empirical data. For example, through model comparisons, we can assess whether response styles may rather be represented by individual profiles (model group 1 in Table 1: independent threshold shifts), or whether there exist systematic components (hence correlations between threshold shifts) between respondents (model groups 2 and 3 in Table 1: constrained or a priori specified response styles). Furthermore, one can make use of the varying degrees of flexibility of the modeling approaches. For instance, one may test whether the symmetry constraint that is applied in multidimensional PCMs (e.g., Bolt & Johnson, 2009; Wetzel & Carstensen, 2017) is reasonable in empirical data, or whether a model using a data-driven approach to estimating the nature of response styles is more appropriate (e.g., Bolt & Johnson, 2009). Finally, one may test whether response style factors have a differential impact on single items through discrimination parameters (Falk & Cai, 2016), and whether one

can explain this influence through, for example, item attributes (see Henninger & Meiser, 2019). Such model extensions add to the modeling framework and may serve to testing specific research questions on response styles.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561–573. doi: 10.1007/BF02293814

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences*, *40*, 1235–1245. doi: 10.1016/j.paid.2005.10.018

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., & 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

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

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212. doi: 10.1177/0022022100031002003

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 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

Ellingson, J. E., Smith, D. B., & Sacket, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86*, 122–133. doi: 10.1037//0021-9010.86.1.122

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

Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, *17*, 56–69. doi: 10.1037/1040-3590.17.1.56

Henninger, M., & Meiser, T. (2019). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Manuscript submitted for publication.*.

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

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

Kelderman, H. (1996). Multidimensional Rasch models for partial-credit scoring. *Applied Psychological Measurement*, *20*, 155–168. doi: 10.1177/014662169602000205

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`

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in survey. *Applied Cognitive Psychology*, *5*, 213–236. doi: 10.1002/acp.2350050305

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

Meiser, T. (1996). Loglinear rasch models for the analysis of stability and change. *Psychometrika*, *61*, 629–645. doi: 10.1007/BF02294040

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

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100. doi: 10.1177/014662169501900110

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:102447211002

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

Morren, M., Gelissen, J. P., & Vermunt, J. K. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology*, *8*, 159–170. doi: 10.1027/1614-2241/a000048

Muraki, E. (1992). A generalized Partial Credit Model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176. doi: 10.1177/014662169201600206

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

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson,

P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of Personality and Social Psychological Attitudes* (pp. 17–59). San Diego, CA: Academic Press. doi: 10.1016/B978-0-12-590241-0.50006-X

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., & 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., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903. doi: 10.1037/0021-9010.88.5.879

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from `https://www.r-project.org`

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, *4*, 321–333. Retrieved from `http://projecteuclid.org/euclid.bsmsp/1200512895`

Rijmen, F., & De Boeck, P. (2005). A relation between a between-item multidimensional IRT model and the mixture rasch model. *Psychometrika*, *70*, 481–496. doi: 10.1007/s11336-002-1007-7

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

Takane, Y., & de Leeuw, J. (1987, sep). 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

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

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. *Handbook of Statistics*, *26*(06), 643–661. doi: 10.1016/S0169-7161(06)26019-X

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., 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*, 69–81. doi: 10.1027/1614-0001/a000102

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

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

Table 1

Structure of the Different Divide-by-Total Models Accounting for Response Styles

	Response Style Perspective	Response Style Distribution	Response Style Specification	Exemplary Research Question	Linear Predictor	Model Characteristics
(1) Models assuming person-specific threshold shifts that are independent of each other and independent of the primary trait						
Wang et al. (2006): Random Threshold Model	threshold	normal	Response styles as random effects of thresholds that are independent of each other	How can one correct for unknown response styles with little a priori assumptions?	$\theta_n - (\beta_i + \tau_{ik} - \delta_{nk})$	$\Sigma = Diag$
Wang & Wu (2011): Generalized Random Threshold Model	threshold	normal			$\alpha_i(\theta_n - (\beta_i + \tau_{ik} - \delta_{nk}))$	$\Sigma = Diag$; α_i (constant across dimensions)
(2) Models constraining person-specific threshold shifts, but estimating response styles exploratory; response styles are typically independent of the primary trait and other response styles						
Rost (1991), von Davier & Rost (2006): Mixture Distribution Model	threshold	discrete	Constant response styles within latent classes	Illustration of types of response style effects	$\theta_{cn} - (\beta_i + \tau_{ik}) + \delta_{ck}$	response styles are class-specific
Böckenholt & Meiser (2017): Mixture Distribution Model	threshold	discrete	Constant response styles within latent classes; linear relationship of threshold distances between classes	Parsimonious specification of differences in threshold parameters between classes	$\theta_n - (\beta_i + \tau_{ik}) + \delta_{ck}$	contraint for threshold distances for class 2: $a + b(\tau_{1ik} - \tau_{1i(k-1)})$
Moors (2003): Latent Class Factor Analysis	trait	discrete	Discrete, exploratory specification of one response style	What kind of response style is in the data?	$\theta_n - b_{ik} + s_k^{*RS}\theta_n^{RS}$	estimated scoring weights
Bolt & Johnson (2009): Multidimensional NRM	trait	normal	Continuous, exploratory specification of response style(s)	What kind of response style(s) is in the data?	$\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$	estimated scoring weights; $\Sigma = I$
Bolt et al. (2014): Multidimensional NRM	trait	normal	Continuous category preference parameters as response styles	What is the preference of each respondent for each category?	$\theta_n - b_{ik} + \theta_{nk}^{*RS}$	sum-to-zero constraint of category preferences: $\sum_{k=0}^{K} \theta_{nk} = 0$ Σ is estimated

Table 1 continued

Structure of the Different Divide-by-Total Models Accounting for Response Styles

	Response Style Perspective	Response Style Distribution	Response Style Specification	Exemplary Research Question	Linear Predictor	Model Characteristics
(3) Models using a priori specifications of response styles; usually the correlation of response styles to the primary trait and other response styles can be estimated						
Jin & Wang (2014): PCM with threshold dispersion	threshold	lognormal	Weight parameter for person-specific threshold dispersion (ERS)	How large is the degree of ERS?	$\theta_n - (\beta_i + \theta_n^w \tau_{ik})$	thresholds dispersion: $\delta_{nk} = -\tau_{ik}(\theta_n^W - 1)$; $\Sigma = Diag$
Morren et al. (2011): Latent Class Factor Analysis	trait	discrete	Discrete, a priori specified response styles	Does ERS exist in the data?	$\theta_n - b_{ik} + s_k^{*RS}\theta_n^{*RS}$	fixed scoring weights; Σ is estimated
Bolt & Newton (2011), Wetzel & Carstensen (2017), Tutz et al. (2018): Multidimensional NRM / PCM	trait	normal	Continuous, a priori specified response styles; typically symmetric ERS and MRS around the item location	How do response styles correlate with the trait and with each other?	$\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$	fixed scoring weights; Σ is estimated
Falk & Cai (2016): Multidimensional gNRM	trait	normal	Continuous, a priori specified response styles; each item may be impacted by each dimension differently	Which items foster response styles? Which item-level features foster response styles?	$\alpha_i\theta_n - b_{ik} + \sum_{d=1}^{D}(\alpha_{id}^{RS} s_{dk}^{*RS})\theta_{nd}^{RS}$	fixed scoring weights; Σ is estimated; α_{id} for each dimension

Note. NRM: Nominal Response Model; PCM: Partial Credit Model; ERS: Extreme Response Style; MRS: Mid Response Style. We use d for dimensions, n for persons, i for items, k for thresholds, s^* for scoring weights adapted to the threshold probability / logit notation, α for discrimination, θ for person parameters, $b_{ik} = \beta_i + \tau_{ik}$ for item and threshold parameters, δ for person-specific shift in thresholds, the superscript ^{RS} to flag response style traits, Σ for the variance-covariance matrix, $Diag$ to indicate a diagonal matrix, I to indicate an identity matrix.

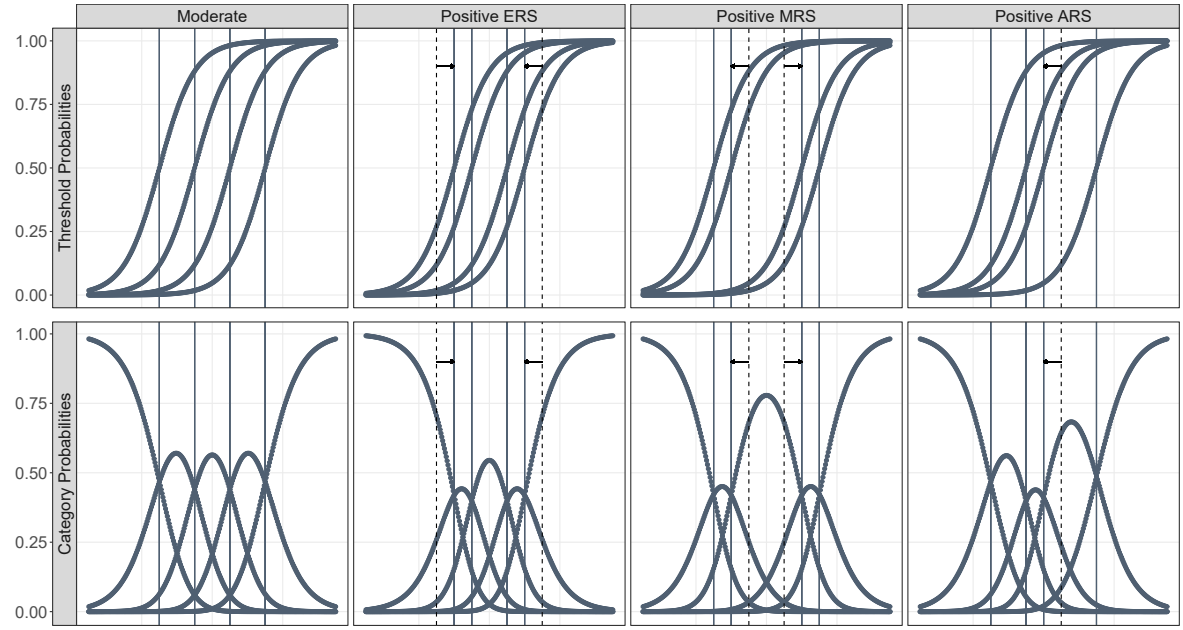

Figure 1. Illustration of threshold (upper row) and category (lower row) probability curves for an item i with five response categories $k \in \{0, ..., 4\}$. From left to right: for moderate respondents, respondents with positive Extreme Response Style (ERS), respondents with positive Mid Response Style (MRS), and respondents with positive Acquiescence Response Style (ARS).

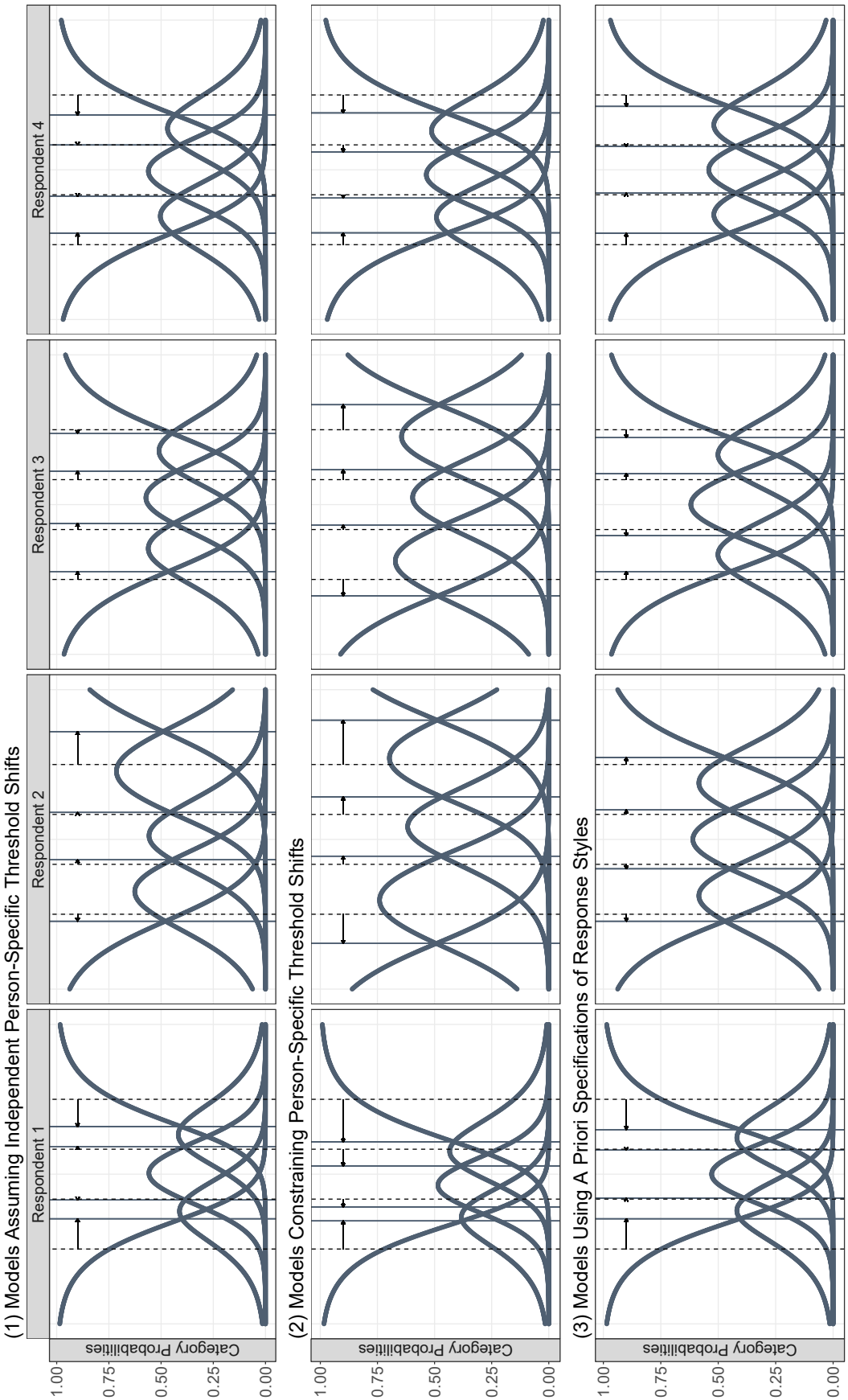

Figure 2. Category probability curves of four exemplary respondents for three groups of IRT models for response styles (see Table 1). Upper row: independent threshold shifts; middle row: threshold shifts are condensed into one dimension; lower row: model with pre-specified ERS and MRS traits.

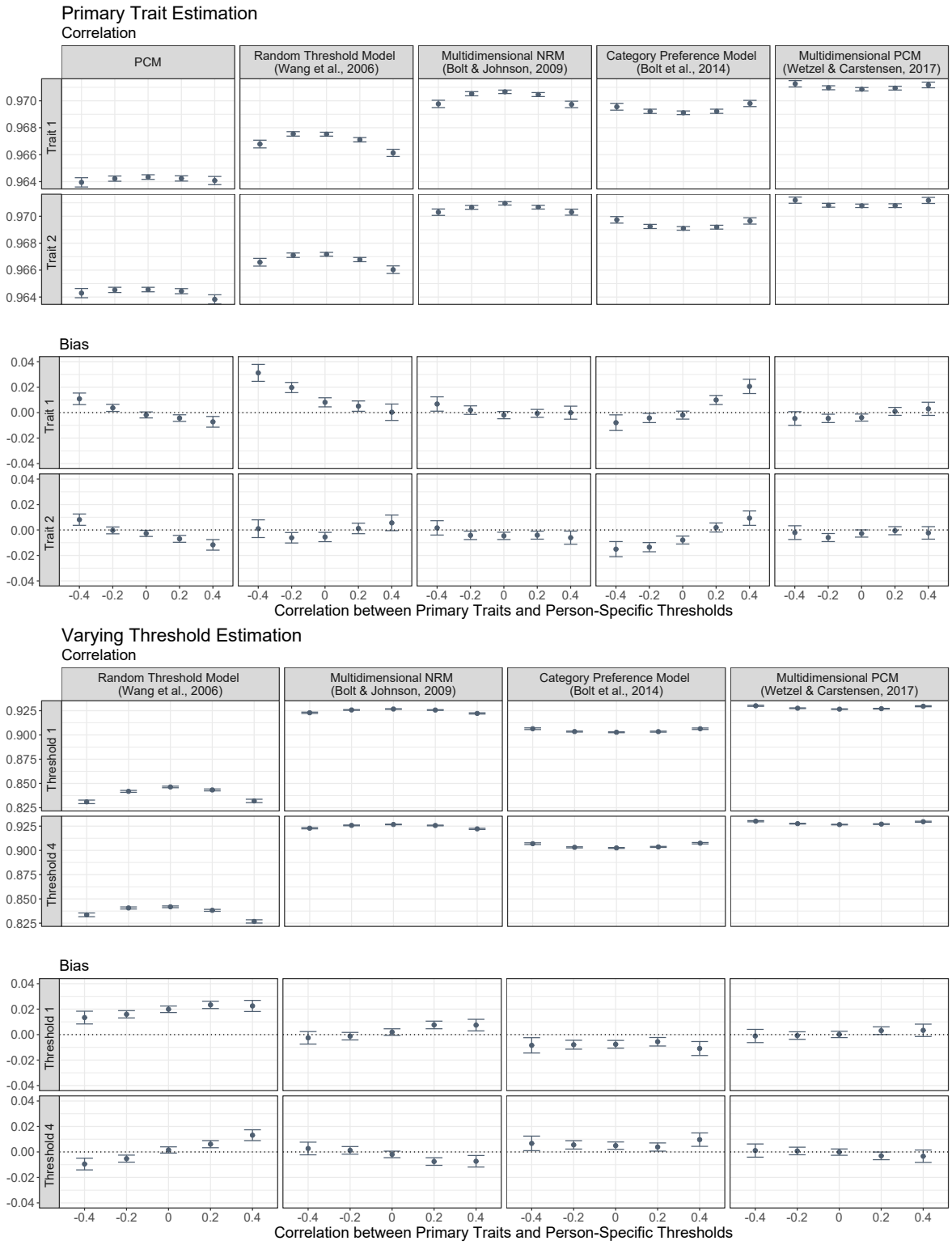

Figure 3. Correlation between true and estimated parameters and mean bias for primary traits (upper panel) and Threshold 1 and 4 (lower panel) in the simulation study for correlations between primary traits and person-specific threshold shifts in the range from $-.5 < \rho < .5$; error bars indicate 95% confidence intervals; PCM: Partial Credit Model.

Appendix A

Linear Parameter Combination Using the Logit Notation for Divide-by-Total Model Variants

Table A1

Linear Parameter Combination Using the Logit Notation ($log\left(\frac{P(X_{ni}=k)}{P(X_{ni}=k-1)}\right)$) for Models Coming From a Threshold-Based Perspective

Models	Original notation	Unified notation	Integrated framework
Wang, Wilson & Shih, 2006	$\theta_n - (\delta_i + \tau_{ij} + \gamma_{nij})$	$\theta_n - (\beta_i + \tau_{ik} - \delta_{nk})$	$\theta_n - b_{ik} + \delta_{nk}$
Wang & Wu, 2011	$\alpha_i(\theta_n - (\delta_i + \tau_{ij} + \gamma_{nj}))$	$\alpha_i(\theta_n - (\beta_i + \tau_{ik} - \delta_{nk})$	$\alpha_i(\theta_n - b_{ik} + \delta_{nk})$
Rost, 1991	$\tau_{vg} + \epsilon_{ixg}$	$\theta_{cn} - b_{cik}$	$\theta_{cn} - b_{ik} + \delta_{ck}$
Jin & Wang, 2014	$\theta_n - (\delta_i + w_n\tau_{ij})$	$\theta_n - (\beta_i + \theta_n^W\tau_{ik})$ $= \theta_n - (\beta_i + \tau_{ik}) - \tau_{ik}(\theta_n^W - 1))$	$\theta_n - b_{ik} + \delta_{nik}$ with $\delta_{nik} = -\tau_{ik}(\theta_n^W - 1)$

Note. In the unified notation and the integrated framework, we use n for persons, c for latent subpopulations, i for items, k for thresholds with $k \in \{1, ..., K\}$, α for discrimination, θ for person parameters, $b_{ik} = \beta_i + \tau_{ik}$ for item and threshold parameters, δ for person-specific shift in thresholds.

Table A2

Linear Parameter Combination Using the Logit Notation ($log\left(\frac{P(X_{ni}=k)}{P(X_{ni}=k-1)}\right)$) for Models Coming From a Trait-Based Perspective

Models	Original notation	Unified notation	Integrated framework
Moors, 2003; Morren & Vermunt, 2011	$\beta_{0jc} + \beta_{1jc}F_{1i} + \beta_{2jc}F_{2i} + \beta_{3jc}F_{3i}$	$\theta_n - b_{ik} + s_k^{*RS}\theta_n^{RS}$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = s_k^{*RS}\theta_n^{RS}$
Bolt and colleagues, 2009, 2011	$a_{jk1}\theta_1 + ... + a_{jkD}\theta_D + c_{jk}$	$\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$
Bolt, Lu & Kim, 2014	$a_{ik}\theta_r + w_{rk} + c_{ik}$,	$\theta_n - b_{ik} + \theta_{nk}^{*RS}$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = \theta_{nk} - \theta_{n(k-1)}$ for $k \in \{1, ..., K\}$
Wetzel & Carstensen, 2017[a]	$\sum_{q=1}^{S} w_{qiy}\theta_{jq} - \delta_{iy}$	$\theta_n - b_{ik} + \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = \sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$
Tutz, Schauberger & Berger, 2018	$\theta_p + (m - r + 0.5)\gamma_p - \delta_{ir}$	$\theta_n - b_{ik} + ({}^{K}/_{2} - k + 0.5)\theta_n$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = ({}^{K}/_{2} - k + 0.5)\theta_n$
Falk & Cai, 2016	$[\mathbf{a} \circ \mathbf{s}_k]'\mathbf{x} + \mathbf{c}_k$	$\alpha_i\theta_n - b_{ik} + \sum_{d=1}^{D}(\alpha_{id}^{RS}s_{dk}^{*RS})\theta_{nd}^{RS}$	$\alpha_i\theta_n - b_{ik} + \delta_{nik}$ with $\delta_{nik} = \sum_{d=1}^{D}(\alpha_{id}s_{dk}^{*RS})\theta_{nd}^{RS}$

Note. The original notations denote the exponential of the numerator of the category probability notation; in the unified and integrated notation, we use the logit notation for simplification. We use d for dimensions, n for persons, i for items, k for thresholds with $k \in \{1, ..., K\}$, s^* for scoring weights adapted to the logit notation, α for discrimination, θ for person parameters, $b_{ik} = \beta_i + \tau_{ik}$ for item and threshold parameters, δ for person-specific shift in thresholds, and the superscript ^{RS} to flag response style traits; [a]the model formula of Wetzel and Carstensen (2017) can be found in Wetzel (2013).

Appendix B

Exemplary Reformulation of Person-Specific Thresholds Into Scoring Weights

As can be seen in Figure 1, ERS affects the outer thresholds while MRS affects the inner thresholds. ARS affects the threshold separating the middle from the first agreement category, while the threshold probability between the agreement is not affected by ARS (both agreement categories remain equally probable). Table B1 shows threshold probabilities of a model with ERS, MRS, and ARS for an item with $K =$ thresholds.

Table B1

Threshold Probabilities for an IRT Model with ERS, MRS, and ARS

Threshold 1	Threshold 2	Threshold 3	Threshold 4
$\frac{\exp\left(\theta_n - b_{i1} - \delta_{n1}^{ERS}\right)}{1+\exp\left(\theta_n - b_{i1} - \delta_{n1}^{ERS}\right)}$	$\frac{\exp\left(\theta_n - b_{i2} + \delta_{n2}^{MRS}\right)}{1+\exp\left(\theta_n - b_{i2} + \delta_{n2}^{MRS}\right)}$	$\frac{\exp\left(\theta_n - b_{i3} - \delta_{n3}^{MRS} + \delta_{n3}^{ARS}\right)}{1+\exp\left(\theta_n - b_{i3} - \delta_{n3}^{MRS} + \delta_{n3}^{ARS}\right)}$	$\frac{\exp\left(\theta_n - b_{i4} + \delta_{n4}^{ERS}\right)}{1+\exp\left(\theta_n - b_{i4} + \delta_{n4}^{ERS}\right)}$

In case that ERS affects categories 0 and 4 by the same weight, we can restrict $-\delta_{n1}^{ERS} = \delta_{n4}^{ERS}$. The same logic applies to MRS, where the second and third threshold (for $K = 4$) are affected equally by MRS and hence $\delta_{n2}^{MRS} = -\delta_{n3}^{MRS}$. Table B2 shows the resulting category probabilities.

Table B2

Category Probabilities when $-\delta_{n1}^{ERS} = \delta_{n4}^{ERS}$ *and* $\delta_{n2}^{MRS} = -\delta_{n3}^{MRS}$

$$
\begin{aligned}
p(X_{ni} = 0) &= \frac{\exp(0)}{c} \\
p(X_{ni} = 1) &= \frac{\exp(1\cdot\theta_n - b_{i1} - \delta_{n1}^{ERS})}{c} \\
p(X_{ni} = 2) &= \frac{\exp(2\cdot\theta_n - (b_{i1}+b_{i2}) - \delta_{n1}^{ERS} + \delta_{n2}^{MRS})}{c} \\
p(X_{ni} = 3) &= \frac{\exp(3\cdot\theta_n - (b_{i1}+b_{i2}+b_{i3}) - 1\cdot\delta_{n1}^{ERS} + 1\cdot\delta_{n2}^{MRS} - 1\cdot\delta_{n3}^{MRS} + 1\cdot\delta_{n3}^{ARS})}{c} \\
&= \frac{\exp(3\cdot\theta_n - (b_{i1}+b_{i2}+b_{i3}) - 1\cdot\delta_{n1}^{ERS} + 1\cdot\delta_{n3}^{ARS})}{c} \\
p(X_{ni} = 4) &= \frac{\exp(4\cdot\theta_n - (b_{i1}+b_{i2}+b_{i3}+b_{i4}) - 1\cdot\delta_{n1}^{ERS} + 1\cdot\delta_{n2}^{MRS} - 1\cdot\delta_{n3}^{MRS} + 1\cdot\delta_{n3}^{ARS} + 1\cdot\delta_{n4}^{ERS})}{c} \\
&= \frac{\exp(4\cdot\theta_n - (b_{i1}+b_{i2}+b_{i3}+b_{i4}) + 1\cdot\delta_{n3}^{ARS})}{c}
\end{aligned}
$$

Note. c is a normalizing constant with $c = \sum_{j=0}^{K} \exp\left(s_j\theta_n - \sum_{k'=0}^{j} b_{ik'} + \sum_{k'=0}^{j} \delta_{nk'}^{RS}\right)$

Through the weights of the response style parameters, we can see a positive ERS trait decreases the probabilities for categories 1 to 3 ($-\delta_n^{ERS}$), which in a Divide-by-Total model in turn increases the probabilities for categories 0 and 4. A positive MRS trait increases the probability of choosing category 2 (δ_n^{MRS}), and a positive ARS trait increases the probabilities for category 3 and 4 (δ_n^{ARS}). From Table B2 and the consequent person-specific threshold shifts, we can directly derive the scoring weights for ERS $\boldsymbol{s}^{ERS} = (0, -1, -, 1-, 1, 0)$, or alternatively $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$, MRS $\boldsymbol{s}^{MRS} = (0, 0, 1, 0, 0)$, and ARS $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ in a multidimensional PCM (cf. Falk & Cai, 2016; Wetzel, 2013; Wetzel & Carstensen, 2017).

Thus, we can reformulate person-specific threshold shifts into a model formulation based on category probabilities. From the category probability notation, we can derive scoring weights for the respective response style traits.