Dissertation

Psychometric modeling as a tool to investigate heterogeneous response scale use

Mirka Henninger

University of Mannheim

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

July 10, 2019

Supervisors:
Prof. Dr. Thorsten Meiser
Prof. Dr. Eunike Wetzel

Dean of the Faculty of Social Sciences:
Prof. Dr. Michael Diehl

Academic Director of the CDSS:
Prof. Dr. Thomas Gschwend

Thesis Evaluators:
Prof. Dr. Thomas Gschwend
Prof. Dr. Eunike Wetzel

Examination Committee:
Prof. Dr. Thomas Gschwend
Prof. Dr. Thorsten Meiser
Prof. Dr. Eunike Wetzel

Date of Defense:
October 8^{th}, 2019

"A good principle of data analysis is never to fall in love with just one model."

— McCullagh & Nelder, 1983 —

Table of Contents

Acknowledgements

··· Many Thanks To ···

··· Thorsten who shaped my thinking and challenged my work while at the same time giving me the freedom to develop my own ideas and research approaches.

··· Eunike who inspired and guided my thesis through her supervision and her own research.

··· Thomas for accompanying me within the CDSS and teaching me new perspectives on research design and methods.

··· Carolin and her research group for inviting me to Zurich and sharing their thoughts and ideas with me.

··· all CDSS members for their intellectual support and valuable advice.

··· anonymous reviewers, Daniel Bolt, and Esther Ulitzsch whose comments helped to improve my work.

··· Alexander Robitzsch, Hadley Wickham, and many more, for creating wonderful statistical and programming tools.

··· my colleagues Gisela, Daniela, and Christine, Dietrich, Simone, Merle, Jana, Susanne, and Nils for their thorough support and many thoughtful discussions.

··· Hansjörg from whom I learned so much.

··· Franziska for being the best office mate I could imagine.

··· my fellow PhD students and friends Sophie, Lili, Sebastian, and Stella who gave me intellectual and moral support and shaped my journey in the last years.

··· Luisa, Antje, Jana, Annika, Ricarda, and Martin (Fladerer) who shared my difficult and happy moments throughout this process.

··· Martin who never stopped challenging my work—but always believed in me.

Abstract

Respondents use different ways to respond to rating scale items. Hence, item responses do not only capture the trait to be measured, but also the way respondents react to rating scales. So-called *response styles* have been incorporated in a variety of psychometric modeling approaches and investigated in applied fields. In my dissertation, I address psychometric and substantive research questions with regards to response styles in four research articles.

In the first article, we structure the variety of psychometric approaches accounting for response styles. We propose a superordinate, unifying framework for such models by introducing one common parameterization. This parameterization then guides our analysis of commonalities and differences, assumptions and identification constraints in the psychometric approaches (Henninger & Meiser, 2019a). We build on the proposed framework in our second article. Herein, we highlight application scenarios and demonstrate how assumptions on response styles can be tested through psychometric approaches. We furthermore develop two novel modeling extensions that lift constraints on model parameters or explain the influence of response styles on items through item attributes (Henninger & Meiser, 2019b).

In the third article (Henninger, 2019), I develop a psychometric modeling approach using a theoretically motivated restriction to achieve statistical identification. The model incorporates little a priori assumptions on response styles and retains the flexibility to account for various kinds of response tendencies in the data. Therefore, it is particularly useful in research environments where response styles differ between subgroups of respondents. The new model is tested in a simulation study and illustrated in a multi-country analysis using data measuring the Big Five personality factors.

The fourth article (Henninger & Plieninger, 2019) deals with processes underlying rating scale responses by examining response times. We find that extreme responding follows a different process than agree and mid responding, and that responses that are in line with the response style trait are given faster. Our analyses suggest that every respondent employs some type of response tendencies that facilitate certain category choices in terms of response speed.

In summary, I integrate existing and propose novel psychometric approaches for response style modeling, and provide new insights into the processes impacting rating scale responses. The two perspective on response styles are mutually reinforcing: psychometric models allow us to test assumptions on response styles. In turn, knowledge about the response process guides psychometricians in refining assumptions that are incorporated in modeling approaches.

1 Introduction

This cumulative thesis is based on the following four manuscripts:

Henninger, M., & Meiser, T. (2019a). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Invited Revision Submitted to Psychological Methods*

Henninger, M., & Meiser, T. (2019b). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Invited Revision Submitted to Psychological Methods*

Henninger, M. (2019). A novel varying threshold IRT approach to accounting for response styles. *Manuscript Submitted for Publication to the Journal of Educational Measurement.*

Henninger, M., & Plieninger, H. (2019). Different styles, different times: How response times can inform our knowledge about the response process in rating scales. *Revision Invited by Assessment.*

The focus of the present thesis is the interplay of psychometric modeling approaches and heterogeneous response scale use in psychological measurement. In the synopsis, I therefore highlight the impact of response styles on rating scale measures, present psychometric approaches to account for response styles, and review theoretical foundations of response styles to motivate the research that I have conducted in my dissertation. I then summarize the four manuscripts that form the core part of my thesis. Last, I discuss the findings and their theoretical as well as psychometric implications and open up directions for future research. The four manuscripts are appended to the synopsis.

1.1 Impact of Response Styles on Rating Scale Measures

In the social sciences, researchers often use rating scales to measure latent personality traits, attitudes, or opinions. As they are convenient to apply, familiar to the respondents, and easy to evaluate, rating scales have become a popular assessment tool. However, it has long been known that respondents use the rating scale in different ways: They perceive the category width differently, and therewith differ in their preferences of specific category combinations over others (Berg & Collier, 1953; Couch & Keniston, 1960; Cronbach, 1942; Hamilton, 1968; Hui & Triandis, 1985). Such heterogeneity in rating scale use is called *response styles*. Irrespective of the item's content, some respondents prefer extreme over intermediate categories (*Extreme Response Style*, ERS), or prefer the middle category (*Mid Response Style*, MRS), while others tend to agree to regular as well as reversed-coded items (*Acquiescence Response Styles*, ARS; see Paulhus, 1991; Van Vaerenbergh & Thomas, 2013). Response styles have been shown to be ubiquitous in rating data (e.g., Bolt & Johnson, 2009; Eid & Rauber, 2000; Meiser & Machunsky, 2008; Rost, Carstensen, & von Davier, 1999; Wetzel & Carstensen, 2017). Furthermore, they seem to be stable across content domains and to persist over time (e.g., Van Vaerenbergh & Thomas, 2013; Weijters, Geuens, & Schillewaert, 2010a, 2010b; Wetzel, Carstensen, & Böhnke, 2013; Wetzel, Lüdtke, Zettler, & Böhnke, 2016).

Response styles can impact the measurement of the content trait. When response styles are ignored, they may bias trait estimates, such as cut-offs in diagnostic assessment situations. For example, a response indicating strong agreement to a rating scale item may be the result of a high content trait level; but it may also be the result of a moderate trait level in combination with a tendency to give extreme or acquiescent responses (e.g., Bolt, Lu, & Kim, 2014; Plieninger, 2017; Van Vaerenbergh & Thomas, 2013). Response styles can also influence relations between measured variables, for example correlations between factor scores or facets of content traits (Böckenholt & Meiser, 2017). Last but not least, response styles can bias cross-group comparisons: when different subpopulations have different response styles, comparisons between groups concerning the trait to be measured may be biased (e.g., De Jong, Steenkamp, Fox, & Baumgartner, 2008; Moors, 2004; Morren, Gelissen, & Vermunt, 2012; van Herk, Poortinga, & Verhallen, 2004).

In order to find ways to deal with response styles, researchers have investigated how the measurement process can be altered to reduce response styles, for example,

by varying the number of response categories and labels or using alternative response formats (Böckenholt, 2017; Plieninger, Henninger, & Meiser, 2019; Weijters, Cabooter, & Schillewaert, 2010), but such attempts yielded inconsistent results. Alternatively, psychometric models accounting for response styles in rating data have been developed in the past decades. The latter approach is the main focus of this thesis.

1.2 Psychometric Approaches to Account for Response Styles

Early approaches used simple descriptive statistics as the number of extreme categories chosen per respondents to measure ERS (e.g., Bachman & O'Malley, 1984; Cronbach, 1942; Greenleaf, 1992a, 1992b). Later approaches regressed the content traits on observed ERS scores and used regression residuals for subsequent analyses (e.g., Baumgartner & Steenkamp, 2001; Weijters, Schillewaert, & Geuens, 2008). Various extensions of latent variable models, such as Item Response Theory (IRT) or Structural Equation Models (SEM), were proposed to correct for response styles. For example, mixture distribution models accounted for response styles by allowing item parameters in IRT models to differ between latent subpopulations. Consistently, mixture distribution model analyses have identified one subpopulation with moderate respondents, and one with extreme respondents (e.g., Eid & Rauber, 2000; Meiser & Machunsky, 2008; Moors, 2003; Rost, 1991; Wetzel, 2013). By allowing for variation of item parameters between classes, content trait estimates in each latent class are corrected for response style influences.

When response styles are present, there are two systematic variance components in item responses: content trait variance and response style variance (Baumgartner & Steenkamp, 2001; Wetzel et al., 2013). These variance components can be separated from each other by modeling response styles as additional latent dimensions in variants of IRT models, such as Sequential (Tutz, 1997), Graded Response (GRM, Samejima, 1969) or Divide-by-Total models (e.g., the Partial Credit Model, PCM; see Masters, 1982; Thissen & Steinberg, 1986). In consequence, estimates of respondents' content traits may be corrected for response style influences. In IRT approaches, response styles can be modeled exploratorily and interpreted post hoc (e.g., Bolt & Johnson, 2009; Bolt et al., 2014; Rost, 1991) as well as specified a priori (e.g., Billiet & McClendon, 2000; Böckenholt, 2012; Bolt & Newton, 2011; De Boeck & Partchev, 2012; Jin & Wang, 2014; Thissen-Roe & Thissen, 2013).

How the biasing effects of ERS can be corrected for by incorporating response styles into psychometric modeling approaches is illustrated in Figure 1.1. Data for this example were rating scale responses from $N = 2,112$ respondents to items measuring the construct *Personal Need for Structure* (PNS, here for the facet *response to lack of structure*) from Meiser and Machunsky (2008). In the

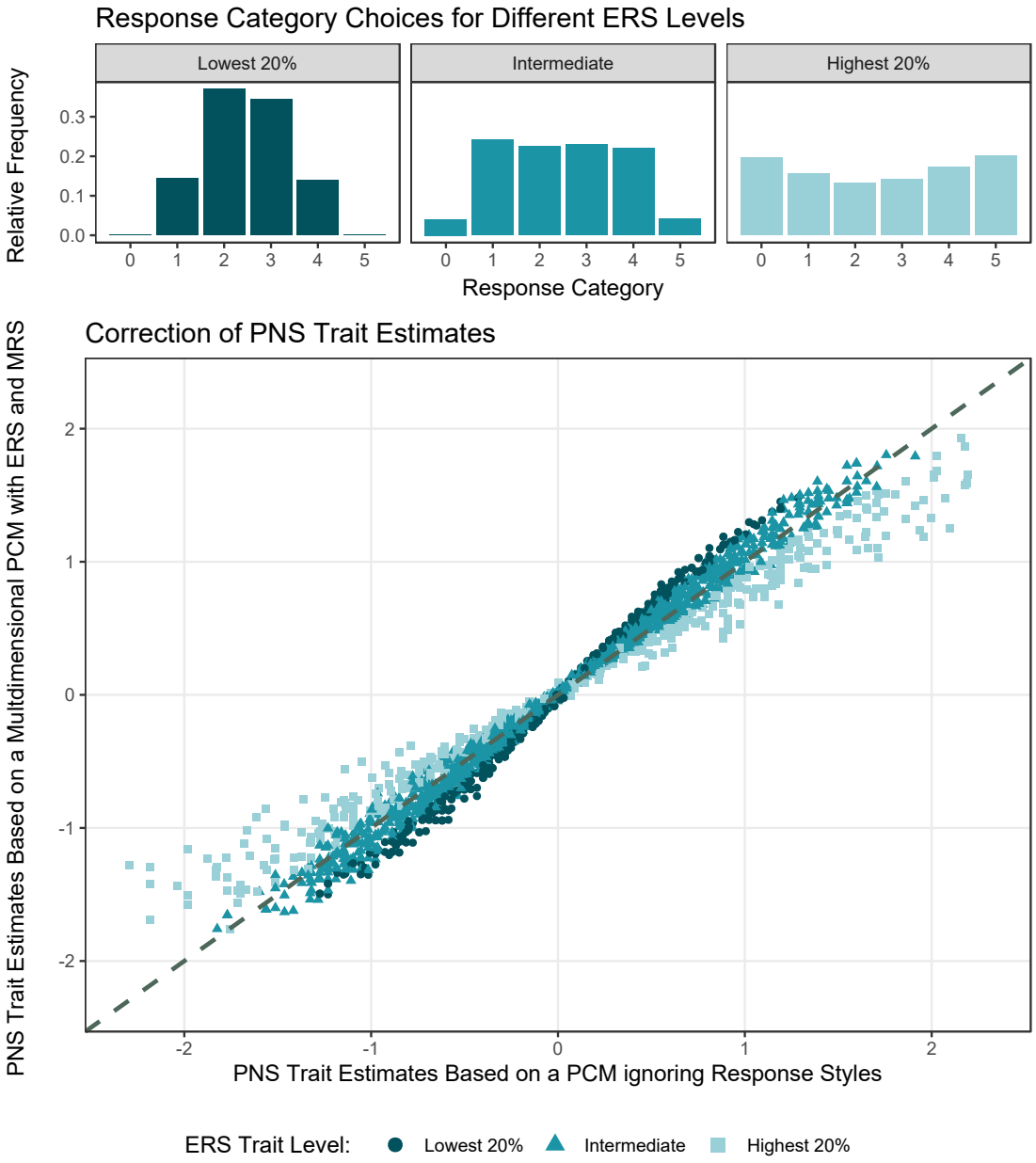

FIGURE 1.1: Upper panel: relative frequency of response category choices for lower and upper 20% quantiles and intermediate levels of Extreme Response Style (ERS); lower panel: correction of Personal Need for Structure (PNS) estimates when ERS is accounted for; content trait estimates for the facet "Response to Lack of Structure" are based on a Partial Credit Model ignoring response styles (PCM; x-axis) and a multidimensional PCM with ERS and Mid Response Style (MRS; y-axis).

upper panel, we see response category choices for different ERS levels (lower and upper 20% quantiles and intermediate levels) based on ERS estimates from a multidimensional PCM. We can see, that in the 20% of the sample with lowest ERS estimates, the extreme categories (here 0 and 5) are never chosen. In contrast, in the highest 20% quantile, the extreme categories are the most frequent ones. For intermediate levels of ERS, choices of the intermediate response categories are more uniformly distributed, with an occasional choice of extreme categories. Hence, we see different ways of using the rating scales between respondents with different ERS levels. In the lower panel, we see the relation of trait estimates of two psychometric models, a PCM ignoring response styles, and a multidimensional PCM that has additional, latent ERS and MRS dimensions. We can see that for high ERS trait levels and low content trait levels, the multidimensional PCM provides an upward correction of content trait estimates (as the "strongly disagree" category is chosen inappropriately often), while they are corrected downwards for high content trait levels (as the "strongly agree" category is chosen inappropriately often), and vice versa for low ERS trait levels. This way, a preference, or avoidance, of extreme categories is accounted for in content trait estimation.

Divide-by-Total Modeling Approaches

In my doctoral thesis, I focus on response style modeling in the Divide-by-Total framework. Therein, one can describe item responses in terms of the threshold probability

$$p(X = k|X \in \{k-1, k\}, \theta, \boldsymbol{b}) = \frac{\exp(\theta_n - b_{ik})}{1 + \exp(\theta_n - b_{ik})} \tag{1.1}$$

that is the conditional probability of choosing either category k or $k-1$ 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. In case that $\theta_n = b_{ik}$, the threshold probability equals .5. Alternatively, we can use a category probability formulation (e.g., a PCM adapted from Masters, 1982) that is defined as a ratio of the exponential of a linear parameter combination divided by its sum across all categories:

$$p(X_{ni} = k) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k} b_{ik'}\right)}{\sum_{j=0}^{K} \exp\left(s_j\theta_n - \sum_{k'=0}^{j} b_{ik'}\right)}. \tag{1.2}$$

The scoring weights s_k describe the relation between trait and category and are usually fixed to $\mathbf{s} = (0, ..., K)$ in a PCM for ordinal rating data. The item-specific category parameter can be decomposed into an item location β_i and an item-specific threshold parameter τ_{ik} with $b_{ik} = \beta_i + \tau_{ik}$ and $\beta_i = (\sum_{k=1}^{K} b_{ik})/K$. The parameter values of the first category are set to 0 ($s_0\theta_n - b_{i0} \equiv 0$). In generalized models, additional item-specific discrimination parameters α_i reflect the influence of the latent trait θ_n on each of the items through $\alpha_i s_k \theta_n - \sum_{k'=0}^{k} b_{ik'}$ (Muraki, 1992).

In Divide-by-Total models, response styles can be incorporated as person-specific shifts in threshold parameters (see Equations 3 and 4 in Chapter 2). These threshold shifts, in consequence, increase the probabilities for certain category combinations while decreasing the probabilities for the others. Figure 1.2 shows category probability curves for one item with five response categories; the vertical lines represent the thresholds. When no response styles are present, thresholds are not shifted (see left column in Figure 1.2). In the presence of ERS, there is a shift of the outer thresholds towards the item location increasing the probabilities for the extreme categories. Similarly, for MRS, the inner thresholds can be shifted outwards making a mid response more probable. For ARS, the threshold separating the middle and agreement categories is shifted so that a response in one of the agreement categories becomes more probable.

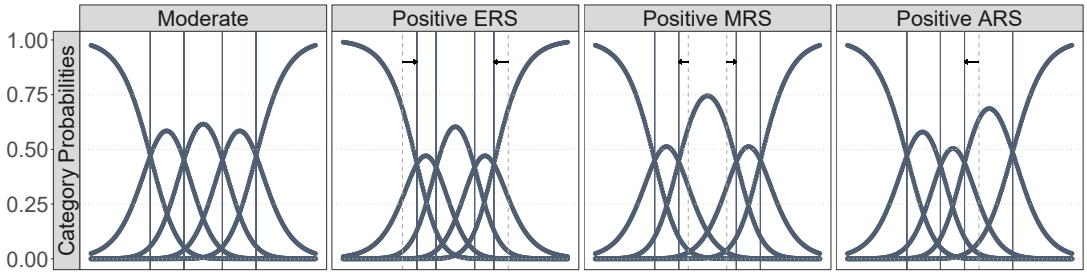

FIGURE 1.2: Illustration of category probability curves for an item with five response categories. From left to right: moderate respondents, respondents with positive Extreme Response Style (ERS), respondents with positive Mid Response Style (MRS), and respondents with positive Acquiescence Response Style (ARS).

Various Assumptions on Response Styles in the Different Models

There exists a variety of modeling approaches for response styles in the IRT literature, and there is no consistent specification of response styles in these models. Rather, response styles are incorporated in many different ways and model-implied

effects of response styles on thresholds substantively vary between the different modeling approaches. For instance, some approaches consider response styles to be variations in item thresholds. These approaches model response styles in terms of random noise due to response heterogeneity (Wang, Wilson, & Shih, 2006; Wang & Wu, 2011) or in terms of threshold dispersion reflecting a combination of extreme and mid responding (Jin & Wang, 2014). Other approaches define response styles through additional response style trait dimensions. Herein, person-specific threshold shifts are composed of response style traits θ_n^{RS} and scoring weights s_k^{RS}. For instance, in case of five response categories scoring weights $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ reflect an ERS dimension wherein the extreme categories become more probably when the ERS trait θ_n^{ERS} is positive. In these multidimensional models, response styles have been incorporated in many different ways: for example, ERS and MRS have been regarded as two separate dimensions, or opposite poles of one dimension (e.g., Falk & Cai, 2016; Jin & Wang, 2014; Thissen-Roe & Thissen, 2013; Tutz, Schauberger, & Berger, 2018; Weijters, Geuens, & Schillewaert, 2010b; Wetzel & Carstensen, 2017). Similarly, different models have used different scoring weights of ERS, MRS, and ARS dimensions (Falk & Cai, 2016; Tutz & Berger, 2016; Weijters, Geuens, & Schillewaert, 2010b; Wetzel & Carstensen, 2017)[1]. Yet other models have incorporated ARS in terms of a shift in item location increasing the probability to agree with the item (e.g., Billiet & McClendon, 2000; Falk & Cai, 2016; Maydeu-Olivares & Coffman, 2006; Wetzel & Carstensen, 2017), or in terms of a mixture process for ARS where agree responses can either be due to acquiescence, or due to content-based agreement (Plieninger & Heck, 2018).

This heterogeneity in modeling approaches illustrates that there are few consistent theoretical assumptions on response styles that are incorporated systematically in the psychometric models. Furthermore, it shows that the way in which response styles influence threshold and category probabilities can vary substantially depending on how response styles are specified in the psychometric model. Wetzel, Böhnke, and Brown (2016) pointed out a lack of model comparisons with regards to the models' ability to control for response styles. Beyond that, there is a need to assess and evaluate the theoretical assumptions on and effects of response styles in the different psychometric models. Therefore, choosing the appropriate model for a specific research question may be demanding, and guidance for model choice and application is missing. Due to the heterogeneity between response style

[1]For example, scoring weights such as $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ for extreme responding, $\boldsymbol{s}^{MRS} = (0, 0, 1, 0, 0)$ for mid responding, $\boldsymbol{s}^{EMRS_1} = (0, 1.5, 2, 1.5, 0)$, or $\boldsymbol{s}^{EMRS_2} = (2, 1, 0, 1, 2)$ for extreme and mid responding, as well as $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ or $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 2)$ for acquiescent responding have been proposed, see Henninger and Meiser (2019a, 2019b).

models, the different approaches are scattered across the psychometric literature, and form several lines of literature that are rarely connected to each other neither holistically examined.

1.3 Theoretical Foundations of Response Styles

In order to incorporate response styles into psychometric modeling approaches in a sensible way, we need to investigate and learn about the nature and underlying processes of response styles.

Response styles have been shown to induce most bias when they are related to the content trait that is aimed to be measured with the rating scale (Plieninger, 2017). Therefore, it is essential to know potential covariates of response styles to evaluate the impact of response styles on measurement validity. However, inconsistent results have been found in terms of the relation of response styles to covariates, such as personality variables, but also gender, age, education, or intelligence (e.g., Böckenholt, 2017; Hamilton, 1968; Meisenberg & Williams, 2008; Moors, 2008; Van Vaerenbergh & Thomas, 2013; Wetzel & Carstensen, 2017). In addition, when covariates are measured with rating scales, their observed values are likely to be confounded with response styles themselves. As first steps towards dissolving the confounds between response tendencies and trait estimates, relations of response tendencies to other traits were assessed through measures by peers (Naemi, Beal, & Payne, 2009) or experimental manipulations of situational factors (e.g., cognitive load, time pressure, rating scale formats; Cabooter, 2010; Kieruj & Moors, 2010; Knowles & Condon, 1999; Weijters, Cabooter, & Schillewaert, 2010); however this led to inconsistent results.

Another important theoretical consideration is whether response styles have an impact on item difficulty. For example, typically extreme or mid responding do not change the item difficulty, as ERS and MRS are considered to be symmetric around the item location (see Figure 1.2). However, one could also hypothesize and test whether, for example, ERS affects the agreement categories more strongly than the disagreement categories which would facilitate agreement to the item for positive ERS traits. In contrast, a tendency to agree with the item irrespective of item content (ARS) is often incorporated in IRT models in terms of a shift on the latent continuum, increasing the probability of agreement categories for respondents with positive ARS levels (Billiet & McClendon, 2000; Falk & Cai, 2016; Maydeu-Olivares & Coffman, 2006; Wetzel & Carstensen, 2017).

Besides, little is known about the cognitive processes underlying response scale use. An early model of such processes has been proposed by Tourangeau and Rasinski (1988). It assumes that respondents who optimize their response read and encode the item content, retrieve relevant knowledge from memory, judge this knowledge, and map their judgment on the rating scale (see also Zaller & Feldman, 1992). In contrast, respondents who use response strategies for at least one of these processes are said be satisficers and to use heuristics such as response styles and invest fewer cognitive resources (Krosnick, 1991). In this vain, ERS has often been associated with low cognitive effort and low motivation (Aichholzer, 2013; Baumgartner & Steenkamp, 2001; Krosnick, 1999). ARS is said to be the result of an intuitive process that leads to spontaneous agreement with the item in contrast to a deliberate process where the item content is evaluated (Knowles & Condon, 1999). Similarly, MRS is regarded to be a result of low cognitive effort: respondents may choose the middle category due to indecision or indifference towards the item content (Baumgartner & Steenkamp, 2001). At the same time, a mid response can be the result of deliberately weighing the pros and cons of the item when a clear-cut decision is not possible (Kulas & Stachowski, 2009). Yet, in the last decades process measures, such as response times, mouse tracking, or eye tracking have become popular in cognitive and experimental psychology (e.g., Franco-Watkins & Johnson, 2011; Heck & Erdfelder, 2016; Hoffman & Rovine, 2007) and ability testing (van der Linden, Klein Entink, & Fox, 2010; van der Linden & van Krimpen-Stoop, 2003) and could inform us how response styles influence the rating process. However, evidence with respect to the relation of process measures and response styles is sparse, mostly inconsistent (Cabooter, 2010; Casey & Tryon, 2001; Hanley, 1965; Knowles & Condon, 1999; Kulas & Stachowski, 2009; Mayerl, 2013; Naemi et al., 2009; Neubauer & Malle, 1997; Swain, Weathers, & Niedrich, 2008), and focuses on data quality, but not on understanding processes underlying rating scale use or response styles themselves.

1.4 The Present Research

With this thesis, I examine how response styles are incorporated into psychometric measurement models, extend the proposed models, and provide insights into the response process.

In two manuscripts (Henninger & Meiser, 2019a, 2019b), we highlight commonalities and differences of different psychometric model from the Divide-by-Total model family. We make the models' assumptions on response styles explicit by

integrating them into one superordinate framework. Therewith, we can regard and examine the psychometric response style literature holistically, assess the ability of modeling approaches to estimate and account for response styles, and extend existing approaches by new models with specific theoretical assumptions.

In the third manuscript (Henninger, 2019), I propose an approach to modeling response styles that incorporates theoretically motivated assumptions on heterogeneous response scale use. Through a new identification constraint, response styles can be reflected by model parameters in a flexible way. At the same time, the constraint allows us to account for response styles such as ERS or MRS that are typically encountered in rating data. The model is particularly useful in research scenarios where little is known about the type of response style in the data or where response styles may differ between sub-groups of respondents as is the case in cross-cultural research settings.

To increase knowledge on response styles and the response process, the fourth manuscript (Henninger & Plieninger, 2019) aims to uncover the cognitive processes underlying heterogeneous response scale use. We use response times to examine how response styles influence the choice of category combinations at the level of single responses, at the level of respondents, and their interactions.

These four manuscripts increase our knowledge on psychometric modeling of response styles, as models are jointly assessed, compared and new model extensions are developed. What is more the manuscripts also increase our knowledge on response styles themselves through shedding light onto the processes underlying rating scale use. The interplay of both aspects add to the response style literature: psychometric models increase knowledge on the nature of response styles and response processes, and at the same time insights into these processes inform and improve psychometric measurement.

2 Different Approaches to Modeling Response Styles in Divide-by-Total IRT Models

Henninger, M., & Meiser, T. (2019a). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Invited Revision Submitted to Psychological Methods*

Henninger, M., & Meiser, T. (2019b). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Invited Revision Submitted to Psychological Methods*

In two manuscripts, we examine the variety of psychometric modeling approaches accounting for response styles. As the models parameterize response styles in different ways, model-implied assumptions on response styles and how they affect threshold and category probabilities are difficult to assess. The heterogeneity between modeling approaches complicates selecting the modeling variant that is most appropriate to correct for or measure response styles in a specific research setting. Therefore, we integrate different modeling approaches for response styles from Divide-by-Total models into one superordinate framework. We propose a common formulation for response styles making assumptions and implications of response style parameterization explicit. We then highlight applications and implications that arise from the joint framework and extend it by proposing new response style model variants.

2.1 Integrating two Lines of Literature Into one Superordinate Framework

In the psychometric literature, there are two perspectives on response styles that have formed two separate lines of literature. One line of literature regards response styles as heterogeneity in item thresholds. In consequence, these models allow threshold parameters to differ between respondents or subpopulations of respondents. For example a shift in the upper and lower thresholds towards the item location increases the probability of choosing one of the extreme categories (see Figure 1.2; Böckenholt & Meiser, 2017; Jin & Wang, 2014; Rost, 1991; Wang et al., 2006; Wang & Wu, 2011). Another line of literature parameterizes response styles as additional person traits. For example, a respondent with a positive ERS trait has a higher probability to choose an extreme category than a respondent with the same content trait level, but medium or negative ERS trait (Bolt & Johnson, 2009; Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Moors, 2003; Morren, Gelissen, & Vermunt, 2011; Wetzel & Carstensen, 2017). The perspective on response styles is closely related to the use of a threshold or category probability notation: in models incorporating response styles as heterogeneity in thresholds, usually a threshold probability formulation (e.g., Equation 1.1) is chosen. In contrast, when response styles are modeled as additional person traits, a category probability formulation (e.g., Equation 1.2) is used.

In order to integrate models with a threshold and trait perspective on response styles, we propose a joint model formulation: we parameterize response styles as person-specific shifts in threshold parameters δ_{nk} for person n and threshold k. This parameterization combines the two lines of literature and allows us to regard response styles in terms of threshold and category probabilities for K thresholds and $K+1$ response categories ($k \in \{0, ..., K\}$):

$$p(X = k | X \in \{k-1, k\}, \theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(\theta_n - b_{ik} + \delta_{nk}\right)}{1 + \exp\left(\theta_n - b_{ik} + \delta_{nk}\right)} \tag{2.1}$$

and

$$p(X_{ni} = k | \theta, \boldsymbol{b}, \boldsymbol{\delta}) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k} b_{ik'} + \sum_{k'=0}^{k} \delta_{nk'}\right)}{\sum_{j=0}^{K} \exp\left(s_j\theta_n - \sum_{k'=0}^{j} b_{ik'} + \sum_{k'=0}^{j} \delta_{nk'}\right)}. \tag{2.2}$$

Herein, θ_n is the respondent's trait parameter, b_{ik} is the item-specific category

parameter for item i and category k ($b_{ik} = \beta_i + \tau_{ik}$), and δ_{nk} a parameter of a person-specific shift in threshold k with $[\theta, \delta_1, ..., \delta_K]$ following a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$.

2.2 Highlighting Model Assumptions Through a Joint Perspective on Response Styles

We show that the various modeling approaches in the response style literature can be subsumed as special cases under the superordinate framework by either imposing restrictions on δ_{nk}, or $\boldsymbol{\Sigma}$, or both. In our two manuscripts, we distinguish three groups of response style models using different restrictions: approaches assuming response styles to be random noise, approaches modeling response styles exploratorily, and approaches using a priori specified response styles. An example of the first group of models is an approach by Wang and colleagues (Wang et al., 2006; Wang & Wu, 2011) assuming that person-specific threshold shifts δ_{nk} are unrelated to each other and to the content trait(s). Therefore, they restricted the covariance matrix to a diagonal matrix $\boldsymbol{\Sigma} = Diag$. The second group of models account for response styles exploratorily. This group comprises mixture distribution models (Böckenholt & Meiser, 2017; Moors, 2003; Rost, 1991) and multidimensional extensions of the Nominal Response Model (NRM; Bolt & Johnson, 2009; Bolt et al., 2014). In the latter case, person-specific threshold shifts are condensed into additional response style trait dimensions that are modeled exploratorily. For example, Bolt and Johnson (2009) added one additional response style trait θ_n^{RS} weighted by estimated scoring weights s_k^{RS} and interpreted it post hoc based on the scoring weights of the response style dimension. The third group of models specifies response styles a priori for example in multidimensional extensions of PCMs. To give an example, ERS can be accounted for by an additional response style trait θ_n^{ERS} that is weighted by a priori fixed scoring weights $\boldsymbol{s} = (1, 0, 0, 0, 1)$ that lead to symmetric, hence negatively correlated, threshold shifts of the outer thresholds (see Figure 1.2; Bolt & Newton, 2011; Falk & Cai, 2016; Jin & Wang, 2014; Morren et al., 2011; Tutz et al., 2018; Wetzel & Carstensen, 2017).

In summary, the proposed framework for Divide-by-Total model extensions for response styles combines two literature lines that have previously parameterized response styles as varying thresholds or additional trait parameters. The framework shows how the different IRT approaches have originally specified response styles and which assumptions on response styles were used to identify the model.

2.3 Applications and Novel Extensions of Response Style Models

In order to illustrate, compare, and extend the different model specifications from the model review and integration, we fit a selection of the models to a standardization sample of the Big Five personality factors (N = 11,724, I = 60, $K+1 = 5$; Borkenau & Ostendorf, 2008). Comparing the modeling approaches, we found an advantage of models specifying response styles a priori (Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017), and of models using item-specific discrimination parameters for Big Five and response style dimensions (Falk & Cai, 2016; Wang & Wu, 2011, see Figure 2.1). These item-specific discrimination parameters reflect the impact of the latent dimensions on items, hence indicate which items are more or less affected by response styles.

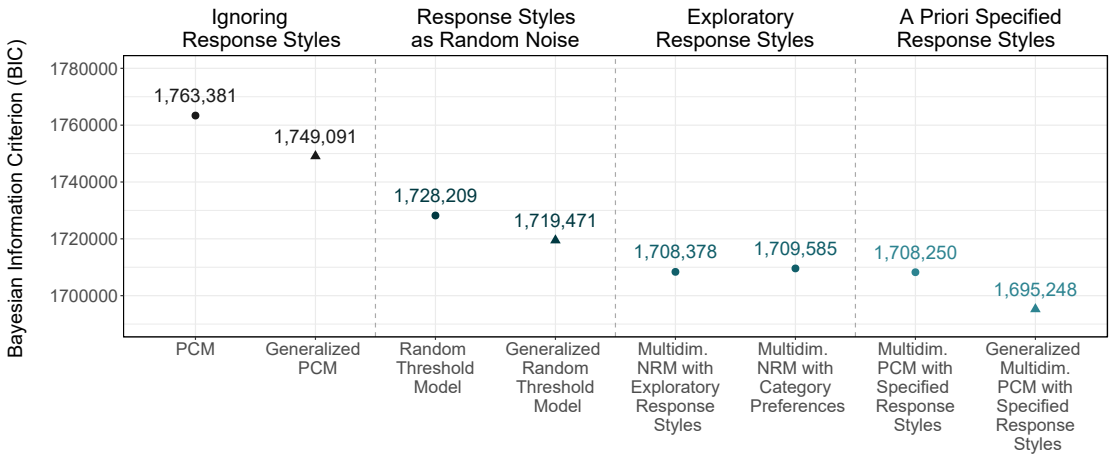

FIGURE 2.1: Overview of the Bayesian Information Criterion (BIC) for the Different Divide-by-Total modeling approaches (see Henninger & Meiser, 2019b); PCM: Partial Credit Model; NRM: Nominal Response Model; the triangular shape indicates a generalized Divide-by-Total model where response style dimensions influence items differently through item-specific discrimination parameters.

Specifying response styles a priori facilitates the interpretation of response style effects and allow us to assess the relations between latent trait and response style dimensions through the variance-covariance matrix $\boldsymbol{\Sigma}$. Furthermore, estimated discrimination parameters inform us to what extent single items are affected by latent response style dimensions. Yet, specifying response styles a priori and estimating differential influence of the response style dimensions on items both come with drawbacks. In the former case, assumptions on the type and nature of response styles must be made. Such assumptions may be that, for example, threshold shifts for ERS are symmetric around the item location, or that certain

thresholds are affected or unaffected by specific response styles (see Figure 1.2). In the latter case, a high number of additional parameter must be estimated, as item-specific discrimination parameters are introduced for each content trait and response style dimension. Therefore, we extend the modeling framework by two novel approaches: First, we lift equality constraints from scoring weights, and second we inform the estimation of discrimination parameters through item attributes, such as complexity, negation, and position, to reduce the number of estimated parameters (Henninger & Meiser, 2019b).

In the first model extension, we test whether ARS affects all threshold separating the agreement categories. In multidimensional PCMs, ARS is incorporated through an additional response style dimension θ_n^{ARS} that weighted by category-specific scoring weights ($\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$). This parameterization increases the probability that a respondent with positive ARS traits gives a response in either of the two agreement categories (see Figure 1.2). We proposed to estimate one of the scoring weights instead of fixing it ($\boldsymbol{s}_k = (0, 0, 0, 1, \lambda^{ARS})$ and find that $\lambda^{ARS} = 1.4$, $SE < .01$. Figure 2.2 illustrates the three variants of ARS modeling on threshold shifts and category probabilities. The figure depicts that for $\lambda^{ARS} > 1$, both thresholds of the agreement categories are shifted towards the item location for positive ARS levels.

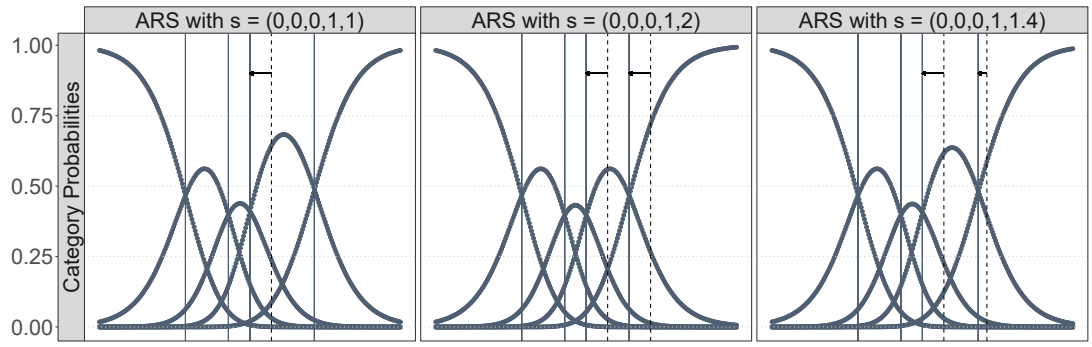

FIGURE 2.2: Category probability curves for three variants of ARS modeling through an adaptation of scoring weights.

In the second model extension, we use indicators of item complexity, item negation, and item position that inform discrimination parameters of response style dimensions. Item attributes can explain some of the influence the response style dimensions have on item responses, still relative model fit indicates that the model cannot account for the all variation in item-discrimination parameters.

All in all, the superordinate framework provides a holistic perspective on psychometric modeling of response styles. It allows us to see and analyze differences between and assumptions of the modeling approaches which facilitates informed

model choice. In turn, addressing research question by comparing specific psychometric response style models increases our knowledge about response styles themselves. For example, the empirical illustration showed that response style dimensions have a differential influence on items and this influence is partly explained by item attributes. Such or similar psychometric models using item attributes as information about response tendencies can be applied in measurement settings to generate, examine, and select questionnaire items.

3 A Novel Varying Threshold IRT Approach to Accounting for Response Styles

Henninger, M. (2019). A novel varying threshold IRT approach to accounting for response styles. *Manuscript Submitted for Publication to the Journal of Educational Measurement.*

The third manuscript (Henninger, 2019) builds upon the insights into psychometric models for response styles from the integrative framework. Herein, we have seen a large variety of ways in which response styles are incorporated into the models and that assumptions made on response styles are rarely made explicit. For instance, Wang et al. (2006) proposed an IRT model that corrects for unknown heterogeneity in response scale usage by specifying content trait and thresholds as random effects. They treat variances in the thresholds as "random noise" (Wang et al., 2006, p. 349) and restrict the variance-covariance matrix to a diagonal matrix for identification. However, the assumption of uncorrelated threshold shifts is likely to be violated in the presence of response styles: ERS and MRS have consistently shown to be present in the empirical data, and these two response styles imply perfect negative correlation of threshold shifts (see Figure 1.2). Therefore, allowing for covariances between threshold shifts is crucial in response style modeling.

At the same time, correcting for unknown heterogeneity in response scale use is highly relevant, for example, in cross-cultural research, where response styles may differ between countries. Ignoring such differences in response styles may lead to biased conclusions drawn from content trait estimates or group comparisons. Hence, accounting for response styles through varying thresholds may be an essential procedure in such research settings. Therefore, I propose a novel varying threshold extension to IRT approaches. The new model is flexible and retains the minimal a priori assumptions of varying threshold models. Besides, it allows for dependencies of varying thresholds that are typically found in empirical data.

3.1 Sum-to-Zero Constraint on Varying Thresholds

Henninger and Meiser (2019a) showed that psychometric modeling approaches for response styles can be parameterized as special cases of the the model proposed in Equations 2.1 or 2.2. For this purpose, restrictions must be imposed either on δ_{nk} or $\boldsymbol{\Sigma}$ to avoid confounds between content traits and response style effects: for example, when all thresholds consistently shift into one direction, the content trait becomes redundant to the varying thresholds and response styles and trait effects cannot be separated (see Henninger & Meiser, 2019a).

The new varying threshold model proposed in this manuscript uses an identification constraint that restricts person-specific threshold shifts to sum to zero across thresholds within persons:

$$\sum_{k=1}^{K} \delta_{nk} = 0 \; \forall \; n. \tag{3.1}$$

Through the sum-to-zero constraint, the model can separate threshold variances δ_{nk} from trait parameters θ_n. Besides, the sum-to-zero constraint introduces dependencies between varying thresholds that are typically found in empirical data, for example in terms of ERS or MRS (e.g., Bolt & Johnson, 2009; Henninger & Meiser, 2019b; Meiser & Machunsky, 2008; Wetzel & Carstensen, 2017).

In the new model, threshold shifts reflect individual respondents' response profiles. These response profiles differ between respondents in terms of their composition of response style effects (e.g., which combination of thresholds are shifted in which direction) and in terms of the magnitude of threshold shifts. Through these individual profiles more unsystematic, person-specific threshold shifts can be captured allowing researchers to account for previously unknown response tendencies. Besides, the sum-to-zero constraint ensures that person-specific shifts in the thresholds reflect the respondent's perception of the rating scale: through the sum-to-zero constraint, person-specific threshold shifts indicate which categories are perceived wider or narrower, and which categories the respondent is more prone to choose. The location of the respondent on the latent continuum, however, is set by his or her content trait and is not affected by response tendencies. Last, through the sum-to-zero constraint dependencies between varying thresholds are implicitly incorporated in contrast to earlier random threshold models (e.g., Wang et al., 2006). Thus, response tendencies that imply symmetric threshold shifts, such as ERS and MRS (see Figure 1.2), can be accounted for by the novel model.

3.2 Relevance in Multi-Group Research Settings

To illustrate the applicability and relevance of the new approach, I conducted a multi-country analysis of four English-speaking countries (Australia, Canada, Great Britain, and USA) using data of a Big Five questionnaire from the Open Source Psychometrics Project (2019). Compared to a PCM, including response styles into the modeling approach improved model fit. However, there were only marginal differences between a random threshold model (e.g., Wang et al., 2006), a multidimensional PCM with ERS and MRS (e.g., Wetzel & Carstensen, 2017), and the novel model using a sum-to-zero constraint. An evaluation of variances and correlations between varying thresholds in the new model using a sum-to-zero constraint indicated that ERS was the dominant response style in the data of all four countries. However, also less dominant response tendencies were present and captured by the new model. Figure 3.1 shows response patterns and category probabilities for four exemplary respondents. The leftmost respondent shows a moderate response pattern with little to no shifts in thresholds. The second respondent shows a preference for the extreme categories that is captured by inward shifts in the outer thresholds. The third respondent has a preference for

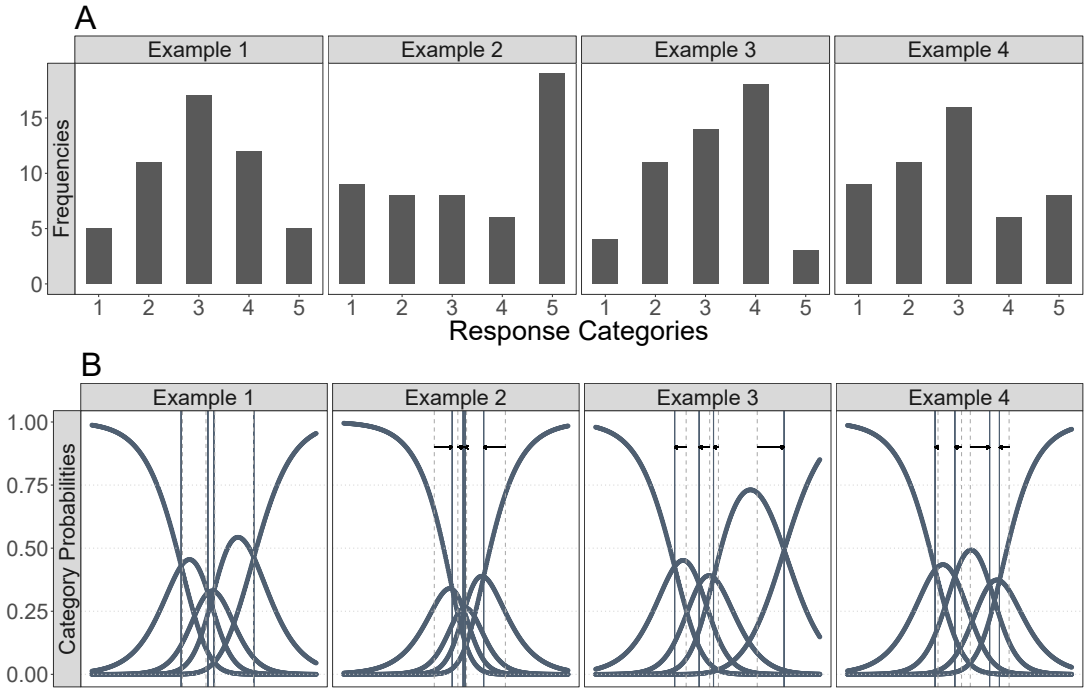

FIGURE 3.1: Category frequencies (A) and threshold shifts (B) for four exemplary respondents; from left to right: respondent with little to no threshold shifts, respondent with ERS, respondent with a preference for the first agreement category, respondent who prefers the middle category, and the highest over the first agreement category.

the first agreement category that is reflected by outwards shifts of the thresholds bounding this category. In the rightmost column, a respondent with a preference for responses in the middle category and in the highest over the first agreement category is shown.

Even though the empirical differences between response style models were marginal, the analysis showed that, besides an extreme response tendency, initially unknown, response tendencies were present in the data. These less dominant and unspecified response tendencies can be captured and described by the model using a sum-to-zero constraint on varying thresholds.

The proposed model extends the bouquet of psychometric approaches by a theoretically motivated IRT variant that explicitly defines how assumptions on heterogeneous response scale use are translated into model parameters. The novel approach can control for previously unmodeled response styles in psychological measurement and is thus well suited for contexts in which the specific response tendencies are unknown. In addition, it adds to the toolbox of approaches investigating response styles as a psychological phenomenon. Herein, it has the potential to become a valuable tool to building consistent theories about heterogeneous response scale use.

4 Different Styles, Different Times: How Response Times can Inform our Knowledge About the Response Process in Rating Scale Measurement

Henninger, M., & Plieninger, H. (2019). Different styles, different times: How response times can inform our knowledge about the response process in rating scales. *Revision Invited by Assessment.*

Examining the literature on psychometric modeling approaches for response styles, we learned about the heterogeneous ways in which response styles are incorporated in the different IRT modeling approaches (Henninger, 2019; Henninger & Meiser, 2019a, 2019b), but the examination also demonstrated that few consistent theoretical assumptions exist about response styles themselves. It is essential to gain more insights into the processes underlying rating scale responses in order to base assumptions on response styles in psychometric models on theoretical grounds. Considering process measures, such as an analysis of response times, may be a means to this end (Fekken & Holden, 1994). Since, item responses are not only an observable representation of the latent content trait, but also of response styles, response times should indicate processes related to both content trait and response tendencies. Therefore, response times can be used to evaluate the often made claim that response styles arise from reduced cognitive effort of the respondent (Aichholzer, 2013; Krosnick & Presser, 2010), and to inform psychometric measurement of content traits and response styles.

4.1 Investigating Three Types of Effects on Response Times

We investigated three different types of effects that response styles can have on response times. Response times may differ between responses of a certain type (e.g., extreme vs. non-extreme responses), between respondents with specific response style trait levels (e.g., respondents with high or low ERS trait levels), and these effects may interact (e.g., a response that is in line with the response style trait may be faster).

We specified a multilevel modeling approach to predict individual log response times of person n and item i using item responses (e.g., $X_{in}^{Extreme}$) on Level 1, respondents' response styles (e.g., θ_n^{ERS}) on Level 2, and their cross-level interaction (e.g., $\theta_n^{ERS} X_{in}^{Extreme}$) as predictor variables. We used dichotomous, dummy-coded indicators for the type of responses (e.g., $X_{in}^{Extreme}$). Thus, in case of extreme response type, extreme responses were coded 1, while intermediate categories were coded 0. For agree responses, agreement categories were coded 1, and for mid responses, the middle category (if applicable) was coded 1. To form latent response style traits (e.g., θ_n^{ERS}), we use a latent aggregation procedure (Lüdtke et al., 2008) implemented in Mplus 7.4 (Muthén & Muthén, 2012) that takes sampling error into account when Level 1 variables are combined to form Level 2 variables.

We used effect-coded item fixed effects ($\sum_{i=2}^{I} \beta_i X_i^{item}$) using X_1^{item} as a reference to account for response time differences due to item attributes. Furthermore, we allowed for random intercept parameters to account for differences between respondents in their response time levels and random slope parameters to examine cross-level interaction effects between response style traits and item responses. Hence, the joint model is given by

$$
\begin{aligned}
&\log \text{Response Time}_{in} = \\
&\gamma_{00} + && \text{Grand Mean} \\
&\sum_{i=2}^{I} \beta_i X_i^{item} + && \text{Item Effects} \\
&\gamma_{10} X_{in}^{Extreme} + \gamma_{20} X_{in}^{Agree} + \gamma_{30} X_{in}^{Mid} + && \text{Level 1: Response} \\
&\gamma_{01} \theta_n^{ERS} + \gamma_{02} \theta_n^{ARS} + \gamma_{03} \theta_n^{MRS} + && \text{Level 2: Respondent} \\
&\gamma_{11} \theta_n^{ERS} X_{in}^{Extreme} + \gamma_{21} \theta_n^{ARS} X_{in}^{Agree} + \gamma_{31} \theta_n^{MRS} X_{in}^{Mid} + && \text{Cross-Level Interaction} \\
&u_{0n} + u_{1n} X_{in}^{Extreme} + u_{2n} X_{in}^{Agree} + u_{3n} X_{in}^{Mid} + e_{in} && \text{Variance Components}
\end{aligned}
$$

and captures effects of responses via $\gamma_{10}, \gamma_{20}, \gamma_{30}$, effects of response style trait levels via $\gamma_{01}, \gamma_{02}, \gamma_{03}$, and cross-level interaction effects via $\gamma_{11}, \gamma_{21}, \gamma_{31}$.

4.2 Response Styles Facilitate Choices of Certain Categories

We applied the multilevel model to three datasets with different characteristics (different sample sizes, different number of response categories, different levels of heterogeneity between items; Fladerer & Misterek, 2018; Pfister, 2018; Plieninger et al., 2019) that contained response times for each item response. Across studies, we found consistent results.

On the response level, response times increased for agree and mid responses, indicating that agree and mid responses might be related to cognitive burden and to be a deliberate process. On the level of the respondent, we found that the higher the ERS trait, the slower was the response. This result is contrary to the claim ERS is associated with low cognitive effort (e.g. Aichholzer, 2013; Baumgartner & Steenkamp, 2001; Krosnick, 1999). In all datasets and across all response styles, we found negative cross-level interaction effects of item responses and response style traits on response times. So respondents were faster when they gave a response that matched their response style trait. Substantively spoken, following the response style trait facilitated the choice of the related response categories in terms of response speed.

We can gain further insights into the cognitive processes when examining the cross-level interaction effect through model-based prediction lines (Figure 4.1, here for Study 3). The interaction is ordinal for extreme responding, but disordinal for agree and mid responding. Hence, the higher the ERS trait level, the more time did the the respondent take when giving a non-extreme extreme response. In

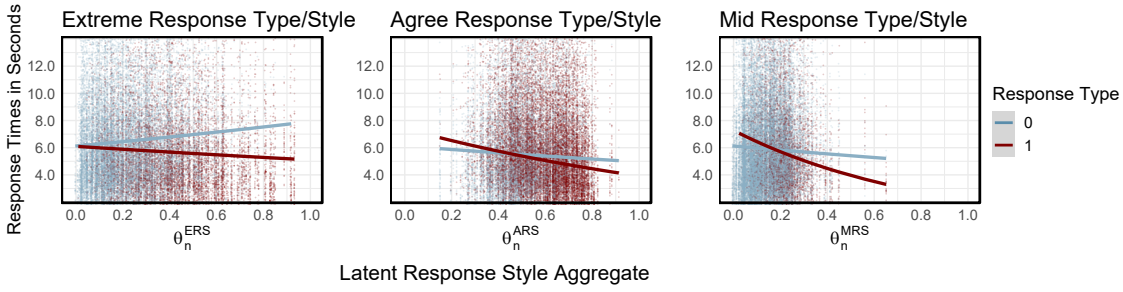

FIGURE 4.1: Scatterplots to illustrate the effect of extreme, acquiescent, and mid responding as a function the respective latent response style trait on response times; exemplary for Study 3.

contrast, for ARS and MRS, agree or mid responses were slower for low response style trait levels, while agree or mid responses were faster for high response style trait levels.

These effects are further illustrated in Figure 4.2 where we show the change in the effect of an item response (e.g., $X_{in}^{Extreme}$) on response times as a function of the latent response style trait (e.g., θ_n^{ERS}). For example, we see that the higher the ERS trait, the more response times decreased when an extreme response was given, compared to a non-extreme response. For ARS and MRS, we again see the disordinal interaction effect, as giving an agree or mid compared to a non-agree or directed response increased response times for low trait levels (positive conditional effect), but decreased response times for high trait levels (negative conditional effect). The vertical lines in Figure 4.2 indicate the boundaries of the regions of significance. Hence, we identified low levels of ERS ($\theta_n^{ERS} < .06$), but intermediate levels of ARS ($0.39 < \theta_n^{ARS} > 0.49$) and MRS ($0.15 < \theta_n^{MRS} < 0.20$) as neutral areas where the conditional effect is not significantly different from 0 and respondents are neither faster nor slower when they give a certain type of response.

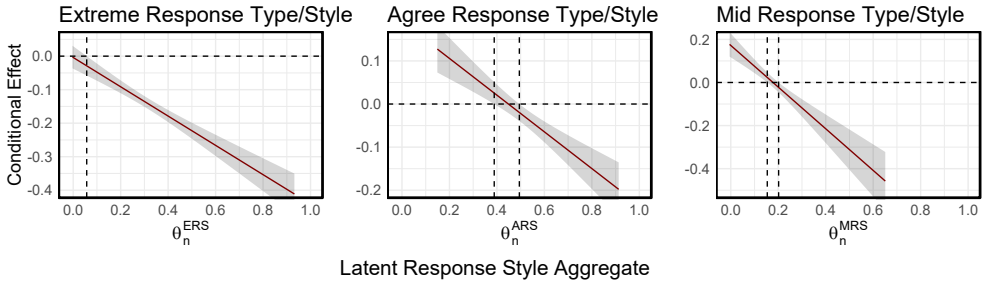

FIGURE 4.2: Conditional effect of giving an extreme, acquiescent, and mid response as a function of the respective latent response style trait on response times; exemplary for the dataset by Fladerer & Misterek (2018).

4.3 Learning About Response Styles from Process Measures

Our results shed light onto the cognitive processes underlying response styles. We showed that only at very low ERS trait levels, giving an extreme response did not influence response times. However, respondents with slightly moderate to high ERS trait levels take more time to respond when they give non-extreme responses. This result suggests that ERS may not necessarily be associated with low motivation or cognitive effort. In contrast to extreme responding, process

patterns of agree and mid responding were very similar. We found that agree and mid responses were slower, and that responses that matched the response style trait were faster (i.e. agree responses were faster for respondents with high levels of ARS, while non-agree responses were faster for respondents with low levels of ARS). These findings suggest a bipolar conceptualization of acquiescence and mid responding where low and high trait levels differentially foster certain response tendencies.

Particularly notable are the highly consistent results across the three datasets which corroborate the effects' robustness and generalizability. Hence, our results are a first step towards making the cognitive processes underlying rating scale use with regards to response styles explicit.

5 General Discussion

In my thesis, I integrated the variety of psychometric modeling approaches accounting for response styles by proposing a joint parameterization in terms of person-specific shifts in threshold parameters. The integration highlights commonalities, differences and assumptions of the different psychometric models. Building on the joint framework, I proposed a new modeling extension that can incorporate a large variety of response tendencies as it allows for dependencies between threshold shifts. The employed sum-to-zero constraint on threshold shifts ensures that response styles do not impact item difficulty, and reflect respondents' perception of category width. To increase our understanding about the mechanisms underlying the response process, I examined the relation of response times and response styles. The results suggest that response styles facilitate the choice of certain categories in terms of response speed and that the process underlying extreme responding is different from agree and mid responding.

5.1 Refining Psychometric Modeling of Response Styles

Uncovering Response Style Parameterizations

The integration of the psychometric modeling approaches (Henninger & Meiser, 2019a) demonstrated that response styles are incorporated in many different ways into the models. Existing models implement response styles as independent random thresholds (Wang et al., 2006; Wang & Wu, 2011), give rise to latent subpopulations (Moors, 2003; Morren et al., 2011; Rost, 1991), account for response styles exploratorily by additional latent dimensions (Bolt & Johnson, 2009; Bolt et al., 2014), or specify them a priori (Bolt & Newton, 2011; Falk & Cai, 2016; Jin & Wang, 2014; Wetzel & Carstensen, 2017). For these different variants, we proposed a common notation, namely parameterizing response styles in terms of person-specific threshold shifts. Based on this parameterization, we made the

model-implied effects on threshold and category probabilities explicit by highlighting restrictions on person-specific thresholds δ_{nk} or the covariance matrix $\mathbf{\Sigma}$ (Henninger & Meiser, 2019a).

We highlighted the heterogeneity in the ways response styles are incorporated in the models, but also commonalities and differences in model assumptions providing guidance for applied researchers. For example, we showed that models cannot account for ERS or MRS when restricting shifts in thresholds to be independent from each other, because ERS and MRS require an inwards or outwards shift of thresholds. In models where ERS and MRS are specified a priori, they are typically constrained to be symmetric around the item location (see Figure 1.2). Besides, ARS is defined as a preference to agree with the item, and thus often implemented as a shift in the threshold separating the non-agreement from the agreement categories (see Figure 1.2). Such a non-symmetric shift leads to a change in item location for respondents with positive (or negative) ARS traits, so the item becomes easier (more difficult; see also Plieninger & Heck, 2018, for a discussion). The joint perspective on psychometric models allows us to investigate, question, and improve modeling assumptions, but also to address more specific research questions about response styles.

Guidance for Informed Model Choice

Psychometric approaches cannot only be tools to correct for response styles in rating data, but also to test specific theoretical assumptions and increase our knowledge about response styles. In the model integration, we illustrated for which research purposes, the different psychometric approaches can be applied (Henninger & Meiser, 2019a, 2019b). For example, in order to control for response styles in different subgroups with unknown response tendencies, a varying threshold approach might be most appropriate (Henninger, 2019; Wang et al., 2006). In order to explore what type of response styles are in the data, a model with the possibility of post hoc interpretations of threshold shift is a useful tool (e.g., Bolt & Johnson, 2009; Bolt et al., 2014; Henninger, 2019). In contrast, if one wants to investigate certain response styles, multidimensional PCMs that allow to explore the relations between content traits and response styles are a sensible choice. For example, one could test with multidimensional PCMs whether ERS and MRS are opposite poles of the same dimension or different dimensions, or assess potential covariates of response styles. Furthermore, estimating discrimination parameters allows us to identify items that are more or less affected by response styles (Falk & Cai, 2016; Henninger & Meiser, 2019b; Wang & Wu, 2011). To the end of providing

guidance to applied researchers, the integration and comparison of the different response style models (Henninger & Meiser, 2019a, 2019b) highlights application scenarios and supports applied researchers to choose a psychometric model that is most appropriate for a specific research question.

A Novel Model with Little A Priori Assumptions on Response Styles

The model integration originates from the need to uncover model-implied assumptions on response styles and effects of response styles on threshold and category probabilities. Due to the different specification, but also parameterizations of response styles (in terms of threshold variations or additional latent traits), and the dispersion of manuscripts across the psychometric literature, these effects were not immediately visible from the original modeling propositions. This demonstrates the need to explicitly discuss assumptions on response styles when developing new modeling extensions in order to justify modeling restrictions, and highlight specific application scenarios for which the present model is more appropriate than competing approaches. I aimed at progressing along this path in proposing a sum-to-zero constraint on varying thresholds (Henninger, 2019).

The novel modeling extension fills a gap in the model structure between the flexible, but theoretically misspecified random threshold models (e.g., Wang et al., 2006) and theoretically sound multidimensional PCMs accounting for ERS and MRS that impose strong restrictions on varying thresholds (e.g., Wetzel & Carstensen, 2017). In this vein, the novel model can account for response styles requiring dependencies between threshold shifts, such as ERS and MRS, but also for more unsystematic response tendencies that can be captured by varying thresholds. Through these characteristics, it is well suited to as an exploratory approach to examine response tendencies, but also to model response style when there is no or little a priori knowledge about their specific types.

Response Process as a Source of Information for Psychometric Modeling

The analysis of response times with regards to response styles (Henninger & Plieninger, 2019) has brought further knowledge on how response styles can be incorporated into psychometric models. For example, the disordinal interaction that we have found for ARS (low ARS levels facilitate non-agree responses, high ARS levels foster agree responses in terms of response speed) speaks in favor of a

response process, where acquiescence and disacquiescence are two opposite poles of one dimension. In consequence, such a process would best be described by a shift (e.g., Maydeu-Olivares & Coffman, 2006; Wetzel & Carstensen, 2017) rather than a mixture model (see Knowles & Condon, 1999; Plieninger & Heck, 2018). In shift models, the ARS trait adds or subtracts to the content trait. Hence, items become easier for high ARS levels, as agreement categories are preferred, and more difficult for low ARS levels, as non-agreement categories are preferred. In contrast, in mixture models agreement can arise due to one of two processes: spontaneous agreement (ARS), and a deliberate process driven by the content trait (Knowles & Condon, 1999). A distinction to shift models is that in mixture models, low levels of acquiescence are not defined as disacquiencence, but as absence of acquiescence (Plieninger & Heck, 2018). However, this assumption would have led to an ordinal, rather than the disordinal interaction effect that we have found for ARS analyzing response times (Henninger & Plieninger, 2019). Therefore, our results speak in favor of incorporating ARS in terms of a shift model using scoring weights $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$, $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 2)$, or similar.

5.2 Contribution to Response Style Theory

Pertinence of Variances and Covariances

This thesis adds to our knowledge about response styles. First, we find substantive correlations between the shifts of the outer and inner thresholds in the data (Henninger, 2019; Henninger & Meiser, 2019a) indicating the presence of ERS. Furthermore, the ERS trait has the largest variance among response style dimensions in all empirical datasets analyzed (PNS dataset in the introduction, Big Five standardization sample, Big Five IPIP sample, and in all three datasets in the response time analyses). These results indicate that mainly extreme responding drives the response process. Furthermore, correlations between content traits and response styles seem to be present in empirical data and crucial in response style modeling. We found medium size correlations between certain content traits and response style dimensions in the empirical datasets (e.g., Henninger, 2019; Henninger & Meiser, 2019b). Plieninger (2017) argued that biasing effects of response styles can be discounted when they are uncorrelated or only weakly related to the content trait. However, the analyses herein show that this is not generally the case and that response styles can be substantially associated with content traits.

Range and Magnitude of the Impact of Response Styles

The analysis of response times as measures of processes showed significant cross-level interaction effects between current item responses and response style traits (Henninger & Plieninger, 2019). This result indicates that response style traits facilitate certain item responses in terms of response speed, when they match the response style trait. We identified response style trait regions for which the effect of giving a response that is in line with the response style trait on response times is significantly positive, negative, or not significantly different from zero. Our analyses show that the range of a neutral area, where response times for both category types (e.g., extreme and non-extreme responses) are equal given a certain level of response style trait (e.g., the ERS trait), is quite small. Consequently, virtually every respondent has a response tendency facilitating certain item responses.

However, the magnitude of the impact of response styles on response times depends on the level of the response style trait. We have shown through our analysis of response times that the impact of extreme responding is only negligible for very low levels of the ERS trait. ERS has an impact on response times for almost all levels of the latent ERS trait insofar that response times increase when non-extreme responses are given. Hence, it seems that it is easy for everyone to give an extreme response, but difficult for respondents with positive ERS traits to give a non-extreme response. However, ARS and MRS follow a disordinal process: while agree and mid responses are faster than non-agree or directed responses for high ARS and MRS levels, they are slower for low ARS and MRS levels. Beyond that, the neutral areas of ARS and MRS are small, so response styles have an impact on response times for nearly all trait levels.

Impact of Response Style Dimensions on Different Items

The comparison of psychometric models has also shown that items are differentially impacted by response style dimensions, as model fit increased when discrimination parameters were estimated (Henninger & Meiser, 2019b). This result indicates that, for example, the ERS dimension has a larger impact on some items than on others. When discrimination parameters are high, the probability of choosing extreme categories substantially increases for positive ERS trait levels, while for negative ERS trait levels, the probability of the intermediate categories increases. When the discrimination parameter is low, the impact of the ERS dimension is small, for positive, negative, and intermediate ERS levels. In addition, we tested whether the impact of response style traits on items can be modeled as a function

of item attributes, such as position, negation, or complexity. We found that indeed such item attributes seem to have an effect on discrimination parameters, indicating that the influence of response styles on item responses is, at least in part, moderated by item attributes.

5.3 Future Directions

A mayor challenge for future research is conceptualizing and examining response styles from a holistic perspective. This perspective should go beyond merely correcting for response styles. Instead, it should consider response styles as a psychological phenomenon on their own. On the one hand, psychometric approaches can be further developed to account for response styles and to increase our knowledge about response styles. On the other hand, insights into the processes underlying response styles can inform the way in which psychometric measurement models specify response styles and help to integrate response styles into a coherent theoretical framework.

In order to account for potential biases due to response styles, psychometric models should be made more accessible to the applied fields. Furthermore, biasing effects of response styles on response times should be examined and taken into account in future developments of psychometric models. In order to learn more about response styles and their influences on rating responses, we need a more coherent picture of response style covariates, knowledge on the moderating influences of item attributes on response style impact, and an understanding of how response styles develop over time, for example in time-intensive assessment situations. These insights will lead to a holistic understanding of response style effects and improve measurement in the social sciences.

Use of Psychometric Models in Applied Research

To ensure that knowledge on and ability to assess and correct for response styles find their way into applied fields, guidance on how to apply psychometric modeling approaches to substantive research questions is essential. Hence, psychometric models for response styles should be made more accessible to applied researchers. As a first step, we provide *R* code of the models that we have illustrated in the first manuscript on Github[1] to make the use of response style models accessible to a wider audience. As a second step, a tutorial on modeling response processes and

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

response styles in the *R* package *TAM* (Kiefer, Robitzsch, & Wu, 2017; R Core Team, 2019) is currently in preparation (Debeer & Henninger, 2019). Further tutorials and open access *R* code are necessary to passing on knowledge about response styles and how to control for them in applied research areas (see also Böckenholt & Meiser, 2017, for a tutorial on modeling subprocesses with IRT).

Response Times as Collateral Information in Psychometric Models

In our analyses of response times, we have shown that response style traits are associated with a relative change in response times when certain categories are chosen. In consequence, using response times as collateral information in IRT models to improve content trait estimation (e.g., Ferrando & Lorenzo-Seva, 2007; Ranger & Ortner, 2011) may be confounded with response style influence. Future research should evaluate to what extent response styles bias parameter estimates in IRT models using response times. For example, one may examine the effects in a simulation study where rating data is generated without response styles, with response styles impacting item responses, and response styles impacting both, item response and response times. Such a study would allow us to evaluate biasing effects of response styles on the estimation of content traits in psychometric models using response times when the impact of response styles on response times is ignored.

Furthermore, knowledge with regard to how response styles influence response times may serve as additional information in IRT models. Ranger and Ortner (2011) proposed the following function to predict log response times:

$$E[log(RT_{in})] = \beta_i^{RT} + \theta_n^{RT} + \alpha_i^{RT} P(X_{ni} = k)$$

where β_i^{RT} and θ_n^{RT} reflect differences between items and persons, respectively, with respect to response times, α_i^{RT} is an item-specific discrimination parameter weighting the impact of $P(X_{ni} = k)$ on response times. The probability of choosing category k ($P(X_{ni} = k)$) is defined as a function of the content trait θ_n^{trait} and an item-category parameter b_{ik} (see Equation 1.2). Thus, the linear predictor informing $P(X_{ni} = k)$ could be extended by response style parameters (e.g., θ_n^{ERS}) to account for the impact of response styles on category probabilities.

Examining Covariates of Response Styles

Even though correlations between response styles and content traits seem to be persistent, they are highly inconsistent across studies (e.g., Austin, Deary, & Egan, 2006; Couch & Keniston, 1960; Grimm & Church, 1999; Hamilton, 1968; He & Van De Vijver, 2013; Moors, 2008; Van Vaerenbergh & Thomas, 2013; Weijters, Cabooter, & Schillewaert, 2010). However, covariates of response styles are usually assessed with self-reports that may be confounded with response styles themselves. Therefore, response style free measurement methods should be used to assess covariates of response styles. In the context of a Bachelor and Master thesis that I supervised (Pfister, 2018; Schreiner, 2019), we examined the relation of ERS, MRS, and ARS and the Big Five personality factors using self-report, peer-report, and implicit measures of personality (Back, Schmukle, & Egloff, 2009; Schmukle, Back, & Egloff, 2008). Our results across measurement methods were mixed, and in part contradict previous findings, or correlations that we have found in the empirical analyses of the Big Five (Henninger, 2019; Henninger & Meiser, 2019b). Future research could extend these analyses to personality measures from the multidimensional forced-choice format (Brown & Maydeu-Olivares, 2011) in order to assess whether more precise predictions on relations between response styles and personality traits can be made. At the same time, including response styles and their correlations to content traits into psychometric modeling approaches appears to be a reasonable strategy to adequately account for response tendencies in rating data (see Plieninger, 2017).

Effect of Item Attributes on Response Style Influence

We have shown that items differ in the strength of impact of the response style dimension and that item discrimination parameters can be informed by item attributes (see Henninger & Meiser, 2019b, see also Meiser, Plieninger, & Henninger, 2019 for using discrimination parameters to examine the influence of latent dimensions on different response subprocesses). This result is based on one analysis of a Big Five standardization sample. However, to assess the coherence of discrimination parameter estimates across studies, the analysis should be carried out in multiple datasets and combined using meta-study techniques. This would lead to a more comprehensive and generalizable picture of item attribute influences on response style use. Knowledge on item attributes that moderate the use of response styles is valuable for item generation and to identify problematic items in test construction.

Response Style Trajectories

There is support for the notion that response styles are consistent across traits and stable over time (Billiet & Davidov, 2008; Danner, Aichholzer, & Rammstedt, 2015; Weijters, Geuens, & Schillewaert, 2010b; Wetzel et al., 2013; Wetzel, Lüdtke, et al., 2016). However, little is known about whether and how response styles rigidify in longer assessment situations and repeated measurements such as panel studies. Do response styles increase, decrease, or stay constant over a longer assessment period? Do extreme responses become faster for ERS respondents at the end of a survey? Can a change point be identified at which response style behavior changes as is the case for careless responding (see Shao, Li, & Cheng, 2016; Yu & Cheng, 2019)? These questions may be answered by modeling person-specific response style trajectories across items using techniques from latent growth-curve modeling (see e.g., Bollen & Curran, 2006; Preacher, Wichman, MacCallum, & Briggs, 2008). For such an analysis data with items in random order is needed to avoid confounding effects of item wording, content, or length on response style or response time influences. Response styles, such as extreme responding, may then be specified as a person-specific varying intercept parameter for ERS and an additional person-specific varying slope parameter for ERS indicating how respondents differ with regards to the changes of ERS impact over the course of the survey.

Generalization of the Notion of Threshold Shifts

We proposed the superordinate framework for response style models with the aim to describe heterogeneity in response scale use in one common notation namely as person-specific threshold shifts. However, the superordinate framework is not limited to psychometric models for response styles, and can be generalized and transferred to other contexts. To give an example, we can use person-specific threshold shifts in order to model dependencies between item responses in ability testing. Here, correct or incorrect responses may inform whether the subsequent item is solved. Similarly, there may be item response dependencies between respondents in case that respondents cheat during testing and copy responses from their neighbors. In personality measurement, such a model would allow us to model consistency or contrast effects in rating scale responses (i.e. whether the same response was given on the previous item, cf. Andrich, Humphry, & Marais, 2012). Thus, the superordinate framework that we proposed in the first manuscript goes beyond providing guidance for model comparison, model choice

and model extensions in the context of response styles. By accounting for potential influences on person-specific threshold shifts, the framework can be extended to other psychometric measurement contexts and allows us to investigate diverse types of influences on dichotomous as well as polytomous item responses.

5.4 Conclusion

In my thesis, I have integrated and extended psychometric modeling approaches, but also provided new insights into the nature and underlying mechanisms of heterogeneous response scale use. One does not work without the other: When we want to learn more about response styles and the response process, psychometric models are essential tools to test competing assumptions. In turn, we cannot develop or refine psychometric modeling approaches without basing assumptions that we incorporate in those models on evidence and sound theory. The interplay of psychometric modeling and knowledge on response styles are the basis for improved and valid psychological measurement.

References

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*, 957–970. doi:10.1016/j.ssresearch.2013.01.002

Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the Rasch model. *Applied Psychological Measurement*, *36*(4), 309–324. doi:10.1177/0146621612441858

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

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*, 491–509. doi:10.1086/268845

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting Actual Behavior From the Explicit and Implicit Self-Concept of Personality. *Journal of Personality and Social Psychology*, *97*(3), 533–548. doi:10.1037/a0016229

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

Berg, I. A., & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement*, *13*, 164–169. doi:10.1177/001316445301300202

Billiet, J. B., & Davidov, E. (2008). Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociological Methods & Research*, *36*, 542–562. doi:10.1177/0049124107313901

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

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

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

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

Bollen, K. A., & Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. New York: John Wiley & Sons.

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

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

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

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

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

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

Casey, M. M., & Tryon, W. W. (2001). Validating a double-press method for computer administration of personality inventory items. *Psychological Assessment*, *13*, 521–530. doi:10.1037/1040-3590.13.4.521

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. doi:10.1037/h0040372

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33*(6), 401–415. doi:10.1037/h0054677

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119–130. doi:10.1016/j.jrp.2015.05.004

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

Debeer, D., & Henninger, M. (2019). Modeling response styles and response processes: A tutorial using the R package TAM. *Manuscript in preparation.*

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

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

Fekken, G. C., & Holden, R. R. (1994). The construct validity of differential response latencies in structured personality tests. *Canadian Journal of Behavioural Science*, *26*, 104–120. doi:10.1037/0008-400X.26.1.104

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543. doi:10.1177/0146621606295197

Fladerer, M., & Misterek, L. (2018). Identity leadership and burnout: A multilevel mediation study. Manuscript in preparation. Retrieved from osf.io/4x9qg

Franco-Watkins, A. M., & Johnson, J. G. (2011). Decision moving window: Using interactive eye tracking to examine decision processes. *Behavior Research Methods*, *43*, 853–63. doi:10.3758/s13428-011-0083-y

Greenleaf, E. A. (1992a). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, *29*(2), 176. doi:10.2307/3172568

Greenleaf, E. A. (1992b). Measuring extreme response style. *Public Opinion Quarterly*, *56*, 328–351. doi:10.1086/269326

Grimm, S., & Church, A. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 415–441. doi:10.1006/jrpe.1999.2256

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

Hanley, C. (1965). Personality item difficulty and acquiescence. *Journal of Applied Psychology*, *49*, 205–208. doi:10.1037/h0022107

He, J., & Van De Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Difference*, *55*, 794–800. doi:10.1016/j.paid.2013.06.017

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Henninger, M. (2019). A novel varying threshold IRT approach to accounting for response styles. *Manuscript Submitted for Publication to the Journal of Educational Measurement*.

Henninger, M., & Meiser, T. (2019a). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Invited Revision Submitted to Psychological Methods*.

Henninger, M., & Meiser, T. (2019b). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Invited Revision Submitted to Psychological Methods*.

Henninger, M., & Plieninger, H. (2019). Different styles, different times: How response times can inform our knowledge about the response process in rating scales. *Revision Invited by Assessment*.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: foundations and illustrative examples. *Behavior Research Methods*, *39*, 101–117. doi:10.3758/BF03192848

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, *49*, 253–260. doi:10.1086/268918

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

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

Kieruj, N. D., & Moors, G. (2010). Variations in response style behavior by response scale format in attitude research. *International Journal of Public Opinion Research*, *22*(3), 320–342. doi:10.1093/ijpor/edq001

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Jornal of Personality and Social Psychology*, *77*, 379–386. doi:10.1037/0022-3514.77.2.379

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

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. Marsden & J. Wright (Eds.), *Handbook of Survey Research* (pp. 263–313). Bingley, UK: Emerald Group Publishing Limited.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*, 489–493. doi:10.1016/j.jrp.2008.12.005

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. doi:10.1037/a0012869

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

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

Mayerl, J. (2013). Response latency measurement in surveys. Detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*, 1–27. doi:10.13094/SMIF-2013-00005

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539–1550. doi:10.1016/j.paid.2008.01.010

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

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

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

perceptions of ethnic discrimination reexamined. *Quality and Quantity*, *37*, 277–302. doi:10.1023/A:1024472110002

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, *20*, 303–320. doi:10.1093/esr/jch026

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, *42*(6), 779–794. doi:10.1007/s11135-006-9067-x

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). Los Angeles, CA: Muthén & Muthén.

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*, 261–286. doi:10.1111/j.1467-6494.2008.00545.x

Neubauer, A. C., & Malle, B. F. (1997). Questionnaire response latencies: Implications for personality assessment and self-schema theory. *European Journal of Psychological Assessment*, *13*, 109–117. doi:10.1027/1015-5759.13.2.109

Open Source Psychometrics Project. (2019). Open psychology data: Raw data from online personality tests. Retrieved from https://openpsychometrics.org/_rawdata/

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). doi:10.1016/B978-0-12-590241-0.50006-X

Pfister, M. (2018). Are extreme and acquiescent response styles related to the implicit self-concept of personality? Unpublished Bachelor Thesis. Mannheim, Germany: University of Mannheim.

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

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

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

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent Growth Curve Modeling.* Thousand Oaks, CA: Sage.

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

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*, 389–406. doi:10.1177/0013164410382895

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

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. / Are the Big Five Rasch scaleable? A reanalysis of the NEO-FFI norm data. *Diagnostica*, *45*(3), 119–127. doi:10.1026//0012-1924.45.3.119

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

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

Schreiner, M. (2019). Personality with style: A multidimensional Item Response Theory approach on the relationship between response styles and personality. Master Thesis in Preparation. Mannheim, Germany: University of Mannheim.

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

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*, 116–131. doi:10.1509/jmkr.45.1.116

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

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

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314. doi:10.1037/0033-2909.103.3.299

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

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

Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the Partial Credit Model. *Applied Psychological Measurement*. doi:10.1177/0146621617748322

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. doi:10.1177/0146621609349800

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265. doi:10.1007/BF02294800

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

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

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., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

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

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science*, *36*(3), 409–422. doi:10.1007/s11747-007-0077-6

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., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (Chap. Res, pp. 349–363). doi:10.1093/med:psych/9780199356942.003.0024

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

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods*. doi:10.1037/met0000212

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, *36*, 579–616. doi:10.2307/2111583

Co-Authors' Statements

Thorsten Meiser

It is hereby confirmed that the following manuscripts included in the present thesis were primarily conceived and written by its first and main author Mirka Henninger.

Henninger, M., & Meiser, T. (2019a). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Invited Revision Submitted to Psychological Methods.*

Henninger, M., & Meiser, T. (2019b). Different approaches to modeling response styles in Divide-by-Total IRT models (Part II): Applications and novel extensions. *Invited Revision Submitted to Psychological Methods.*

Mirka Henninger and Thorsten Meiser developed jointly the idea of the integrating the different IRT models into one superordinate framework. Mirka Henninger planned and carried out the literature review, reformulations to demonstrate model equivalence as well as the analyses of empirical data, new model development and simulation studies. Mirka Henninger was solely responsible for writing the first draft and for finalizing the two manuscripts.

Thorsten Meiser assisted with his knowledge of IRT and response style modeling and ideas for new modeling approaches. Furthermore, he gave helpful comments on draft versions of the manuscript. Apart from that, Thorsten Meiser helped in numerous discussions to refine specific parts of the literature review and empirical analyses.

Thorsten Meiser	Place, Date

Hansjörg Plieninger

It is hereby confirmed that the following manuscript included in the present thesis was primarily conceived and written by its first and main author Mirka Henninger.

Henninger, M., & Plieninger, H. (2019). Different styles, different times: How response times can inform our knowledge about the response process in rating scales. *Revision Invited by Assessment.*

Mirka Henninger developed the idea of the manuscript as well as the design and procedure of the studies. She was solely responsible for all analyses, for writing the first draft and for finalizing the manuscript.

Hansjörg Plieninger contributed in numerous discussions to refine hypotheses and statistical analyses, and gave helpful comments on draft versions of the manuscript.

Hansjörg Plieninger	Place, Date

Different Approaches to Modeling Response Styles in Divide-by-Total IRT Models (Part I): A Model Integration

Mirka Henninger and Thorsten Meiser

University of Mannheim

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.

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 content 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 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 models, a framework that is in commonly used to model and account for response styles. One advantage of Divide-by-Total models is the clear interpretation of thresholds. In Divide-by-Total models, 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. Furthermore, in contrast to GRMs, Divide-by-Total models can accommodate unordered thresholds, which allows capturing very low category probabilities or collapsing categories due to response tendencies. As another advantage, Divide-by-Total models directly reflect the ordinal response process for the trait, whereas IRTree models often dichotomize indicators of the latent trait. In this case, the intensity of category choice (e.g., choosing "strongly agree" instead of "agree") is solely determined through response styles and does not involve the content trait to be measured (although this assumption can be tested, see Jeon & De Boeck, 2016; Meiser, Plieninger, & Henninger, 2019). Divide-by-Total models retain the ordinal response process for the trait and can model response styles as additional trait dimensions or as shifts of thresholds. Finally, Divide-by-Total models allow for exploratory as well as confirmatory analyses of response styles. In IRTree models, in contrast, response processes must be defined a priori and cannot be explored through a data-driven approach. Therefore, 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. In a second article (Henninger & Meiser, 2019), we illustrate the specification and fit of the response style models with a standardization sample of a Big Five inventory (Borkenau & Ostendorf, 2008). Furthermore, we use the integration framework to derive two new model variants that lift certain constraints from model parameters.

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 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).

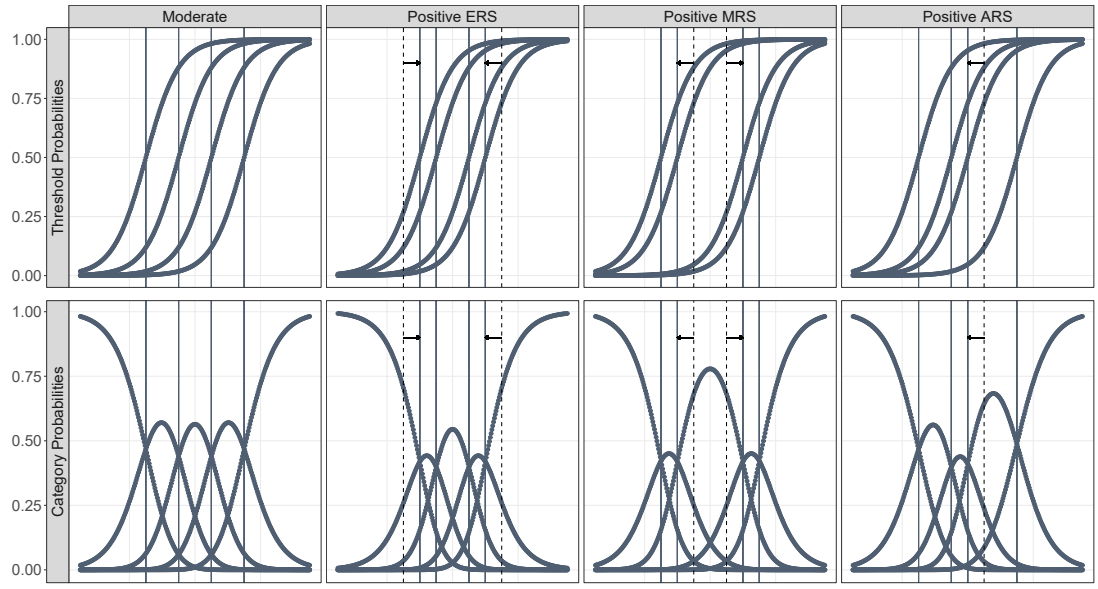

FIGURE 1: Illustration of threshold (upper row) and category (lower row) probability curves for an item i with five response categories $k \in \{0, ..., 4\}$. From left to right: for moderate respondents, respondents with positive Extreme Response Style (ERS), respondents with positive Mid Response Style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

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.

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 $k \in \{0, ..., K\}$[1].

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 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).

[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 content 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).

Of course, the modeling framework in Equations 3 and 4 is not identified as content 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 content 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 content 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 content traits) or anchoring vignettes (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Weijters, 2006; Wetzel & Carstensen, 2017). 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 content 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 content 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

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

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].

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 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

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

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.

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 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. In order to disentangle the content 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 $[\theta, \delta_1, ..., \delta_K]$.

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 structure of threshold shifts across respondents in the data: we see that the profile of threshold shifts is equal

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 content trait						
Wang, Wilson, and Shih (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 and 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 content trait and other response styles						
Rost (1991), von Davier and 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 and 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 and 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, Lu, and Kim (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 content trait and other response styles can be estimated						
Jin and 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, Gelissen, and Vermunt (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 and Newton (2011), Wetzel and Carstensen (2017), Tutz, Schauberger, and Berger (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 and 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.

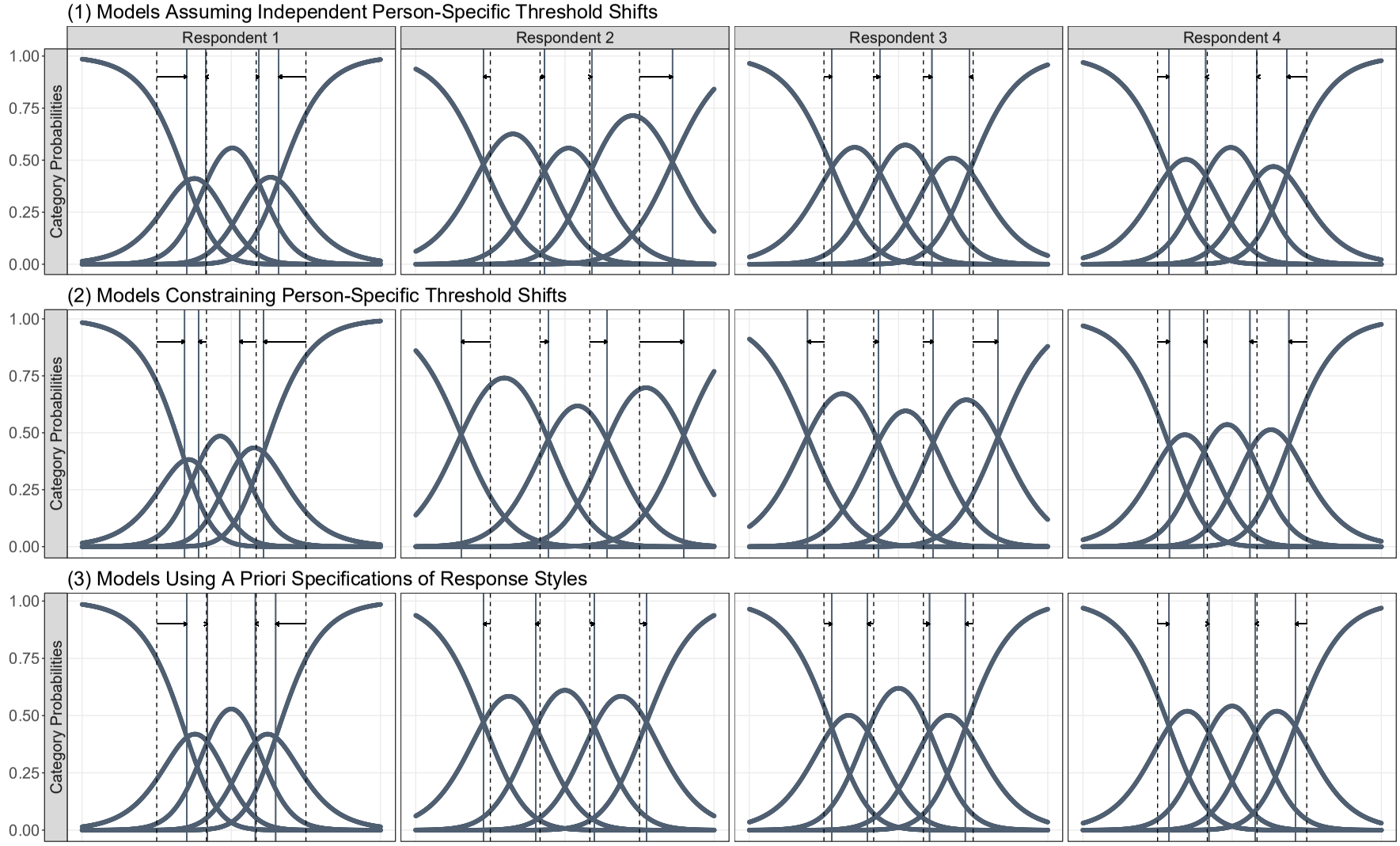

FIGURE 2: Category probability curves of four exemplary respondents for three groups of IRT models for response styles (see Table 1). Upper row: independent threshold shifts; middle row: threshold shifts are condensed into one dimension; lower row: model with pre-specified ERS and MRS traits.

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 can be defined as a linear function of threshold distances in subpopulation 1.

[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.

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 content traits and estimated category scoring weights for one response style dimension freely.

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 $\theta_n - b_{ik}+\sum_{d=1}^{D} s_{dk}^{*RS}\theta_{nd}^{RS}$. 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: For instance, 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 content 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 model 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 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 content 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 $\delta_{nk} = (\frac{K}{2} - k + 0.5)\ \theta_n^{RS}$ (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 content 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 content 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. 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).

[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 $\boldsymbol{s}^{RS} = (0, 1.5, 2, 1.5, 0)$.

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, 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, 2019) 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 $\boldsymbol{\delta}_n = \left(s_1^* \theta_n^{RS}, \quad s_2^* \theta_n^{RS}, \quad \dots, \quad s_K^* \theta_n^{RS}\right)$.

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

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

and the cumulative category scoring weights for the response style trait θ_n^{RS} are given by $\boldsymbol{s} = \left(s_1^*, \quad s_1^* + s_2^*, \quad \ldots, \quad \sum_{k=1}^{K} s_K^*\right) = \left(0, \quad s_1, \quad s_2, \quad \ldots, \quad s_K\right)$. 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 $\boldsymbol{\delta}_n = \left(\theta_{n1}^{RS} - \theta_{n0}^{RS}, \quad \theta_{n2}^{RS} - \theta_{n1}^{RS}, \quad \ldots, \quad \theta_{nK}^{RS} - \theta_{n(K-1)}^{RS}\right)$ with $\theta_{n0}^{RS} \equiv 0$.

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

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

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

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

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 content 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 $\boldsymbol{\delta}_{ni} = \left(-\tau_{i1}(\theta_n^W - 1), \quad -\tau_{i2}(\theta_n^W - 1), \quad \dots, \quad -\tau_{iK}(\theta_n^W - 1)\right)$, and in consequence, the n^{th} row of $\boldsymbol{\delta} \times \boldsymbol{T}^{d}$ is given by

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

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 θ_n^{ERS}, θ_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 $\boldsymbol{\delta}_n = \left(-\theta_n^{ERS},\ \ \theta_n^{MRS},\ \ -\theta_n^{MRS} + \theta_n^{ARS},\ \ \theta_n^{ERS}\right)$.

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

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

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 content trait and response style dimensions.

In the simulation study, we examined content 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 content 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 content dimensions. In order to facilitate estimation of the response style models, each content 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 content and one ERS dimension was fixed to 1, the covariance between the content traits was fixed to $\rho = .2$, and for each replication the correlation between the content traits and the ERS trait was drawn from a Wishart distribution with 5 degrees of freedom and set equal for the two content dimensions. Respondents' trait parameters were generated from a $MVN \sim (\mathbf{0}, \boldsymbol{\Sigma})$.

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 content 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 content 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 $\delta_{nk} = \theta_{nk}^{*RS} = \theta_{nk}^{RS} - \theta_{n(k-1)}^{RS}$. This model can account for the symmetric threshold shifts through ERS and correlations between content traits and ERS. All model were estimated using R (R Core Team, 2019) 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 (Cor $= r(\hat{\theta}_n, \theta_n)$) and mean bias (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 content 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 content 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 content 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 content 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 content and response style traits, we see negative quadratic trends for models restricting the covariance between content 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).

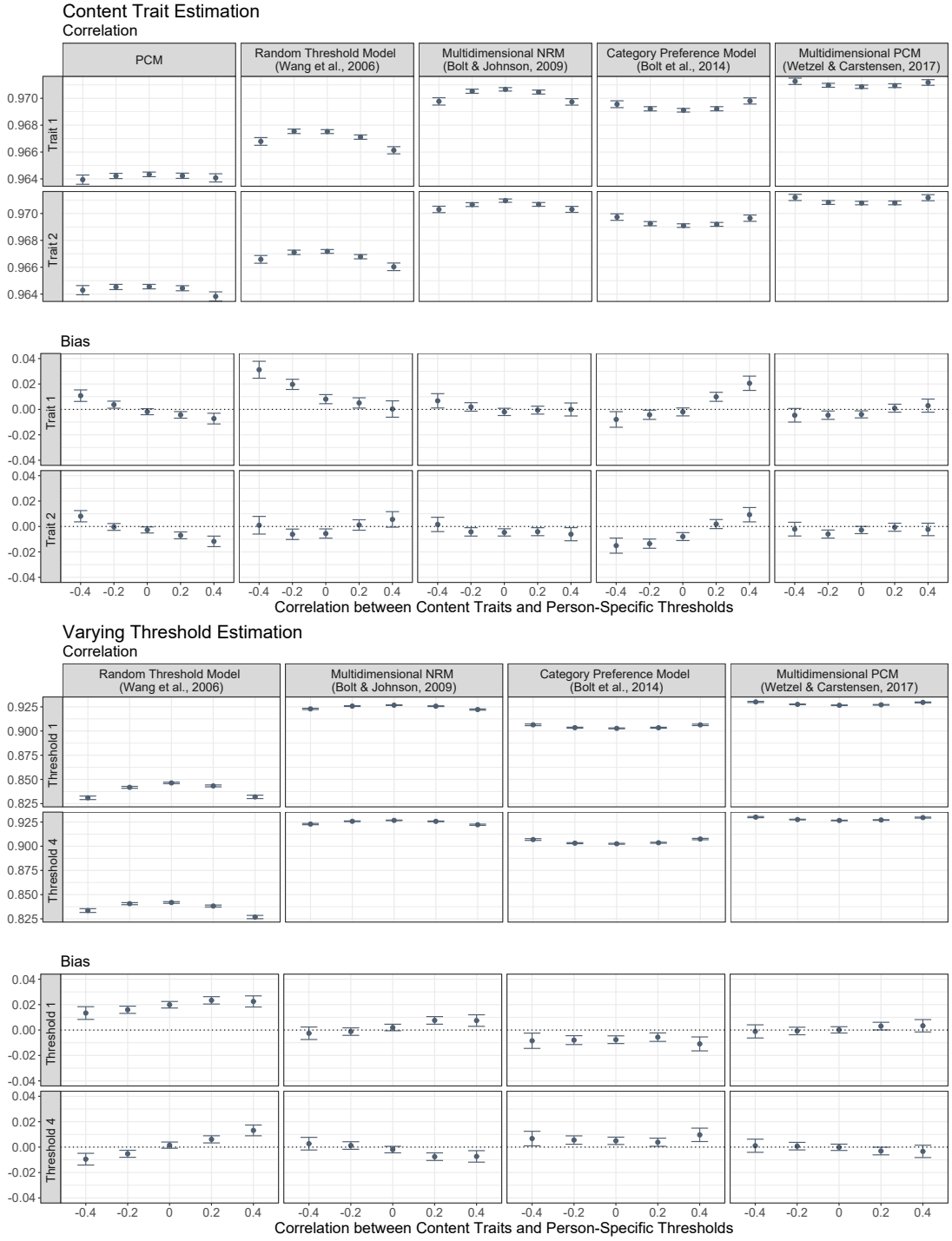

FIGURE 3: Correlation between true and estimated parameters and mean bias for content traits (upper panel) and Threshold 1 and 4 (lower panel) in the simulation study for correlations between content traits and person-specific threshold shifts in the range from $-.5 < \rho < .5$; error bars indicate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

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, content traits were overestimated for negative, and underestimated for positive correlations between content traits and person-specific threshold shifts. For content 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 content 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 content 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. 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, 2019) 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.

In a second article (Henninger & Meiser, 2019), we extend the integrated framework of response style models by applications to empirical data and modeling extensions. We use a standardization sample of a Big Five inventory (Borkenau & Ostendorf, 2008) to illustrate the specification, parameter estimation, and fit of the different response style models. We then propose two modeling extensions that close gaps in the model structure. The first proposition lifts the equality constraint in scoring weights of the ARS dimension and the second proposition increases model parsimony by using item attributes to restrict discrimination parameters of response style dimensions. These novel extensions add to the modeling framework and serve as a guidance to develop new approaches in order to test 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

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

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. *Invited Revision Submitted to Psychological Methods.*

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:1024472110002

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

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). 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). 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. (2019). 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). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408. doi:10.1007/BF02294363

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

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

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. (2006). *Response styles in consumer research* (Doctoral dissertation, Ghent University). Retrieved from https://biblio.ugent.be/publication/4100284/file/4100290

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

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	$\theta_p + (m - r + 0.5)\gamma_p - \delta_{ir}$	$\theta_n - b_{ik} + ({}^{K}/_{2} - k + 0.5)\theta_n^{RS}$	$\theta_n - b_{ik} + \delta_{nk}$ with $\delta_{nk} = ({}^{K}/_{2} - k + 0.5)\theta_n^{RS}$
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.

Different Approaches to Modeling Response Styles in Divide-by-Total IRT Models (Part II): Applications and Novel Extensions

Mirka Henninger and Thorsten Meiser

University of Mannheim

Abstract

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

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

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

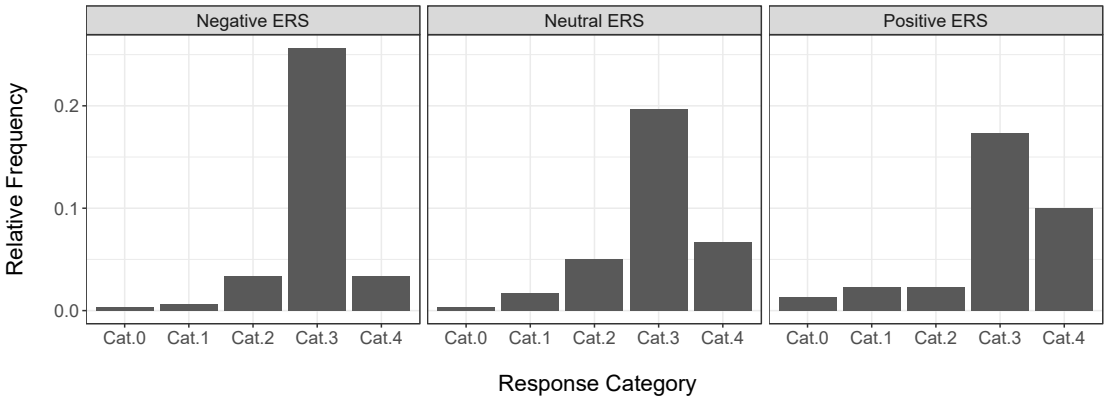

FIGURE 1: Relative frequencies of response category choices for three exemplary respondents based on simulated data with the same manifest mean across items ($\bar{X} = 3$, moderately positive trait levels), but different Extreme Response Style (ERS) levels.

Psychometric Models for Response Styles

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

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

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

and

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

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

Figure 2 shows how such person-specific threshold shifts can be incorporated in IRT models to reflect response styles. The category probability curves are impacted through the inclusion of response styles into the IRT model. For example,

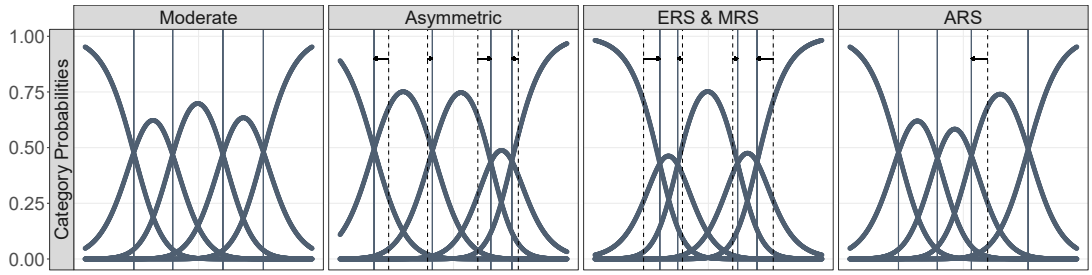

FIGURE 2: Illustration of category probability curves for an item i with five response categories $k \in \{0, ..., 4\}$. From left to right: for moderate respondents, respondents with a unique profile of asymmetric threshold shifts, respondents with positive Extreme and Mid Response Style (ERS & MRS),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

a respondent with asymmetric threshold shifts has a unique profile of response tendencies that leads to, for example, a decrease in probability for the lowest category (column 2). In contrast, ERS is described by a shift of the outer thresholds towards the item location, thereby widening the interval over which the extreme categories have the modal probability. MRS is described by a shift of the inner thresholds away from the item location, thereby widening the interval over which the middle category is most probable. In consequence, the probability of choosing one of the extreme categories or the middle category increases for a person with a given content trait level as a function of ERS and MRS, while the probability of choosing one of the intermediate categories decreases (column 3). For positive ARS levels, the threshold separating the middle from the agreement categories is shifted towards the left, increasing the probability of a response in one of the agreement categories (right column).

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

Table 1, for the composition of δ_{nk} in different response style models see Tables A1 and A2 in Appendix A in Henninger & Meiser, 2019).

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

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

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

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

The third group of models use a priori specifications of response styles. For example, Jin and Wang (2014) assumed that response styles pull apart or push together item thresholds. They introduced a person-specific weight parameter to reflect this dispersion. Other approaches added response style dimensions to a PCM. The scoring weights $\boldsymbol{s}_k$ of response style dimensions are fixed a priori, for example to incorporate ERS, MRS, and ARS traits into the model (see column 3 & 4 in Figure 2 and Table 1). In consequence, correlations between response

style and content trait dimensions can be examined (Bolt & Newton, 2011; Tutz, Schauberger, & Berger, 2018; Wetzel & Carstensen, 2017). Falk and Cai (2016) added item-specific discrimination parameters to describe the impact of response style dimensions on items as a further extension.

Implications of the Integrated Framework and Overview

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

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

In addition, we can use the superordinate framework to derive novel extensions to the existing models. In this vein, we propose two novel model variants that extend existing IRT models for response styles. The first proposition lifts an equality constraint from scoring weights, and in the second proposition, discrimination parameters are specified as functions of item attributes. Both models fit in and extend the model structure, and open up new possibilities to improve measurement of traits and analyses of response styles.

Model Implementation in Standard Software

Subsuming the different Divide-by-Total modeling extensions under the superordinate framework (Equations 1 and 2) allows us to implement the models as

TABLE 1: Exemplary Scoring Weights for an Item With 5 Response Categories

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

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

multidimensional extensions of NRMs (Bock, 1972; Takane & de Leeuw, 1987). As it is not immediately obvious how threshold shifts translate into scoring weights—in particular for models with varying thresholds (e.g., Wang et al., 2006), or for models with category preferences summing up to zero (Bolt et al., 2014)—expressing response style models as multidimensional NRMs allows us to identify the scoring weights that we can use for estimation in standard statistical software (see Henninger & Meiser, 2019). We summarize scoring weights for trait, random thresholds, exploratory response styles, category preferences, and response styles

ERS, MRS, and ARS for an item with $K = 4$ thresholds and $K + 1 = 5$ categories in Table 1.

Standard statistical software programs such as Mplus (Muthén & Muthén, 2012, see also Huggins-Manley & Algina, 2015) or the *R* (R Core Team, 2019) packages *TAM* (Kiefer, Robitzsch, & Wu, 2017) or *mirt* (Chalmers, 2012) have built-in procedures to estimate multidimensional IRT models. Standard software programs implement procedures that allow us to specify whether scoring weights of each item and category for each latent dimension should be estimated, constrained, or fixed to a specific value. For example, we can set up a multidimensional PCM with fixed scoring weights for trait and response style dimensions through specifying that each item relates to both, the content trait and the response style dimensions through the scoring weights from Table 1 (e.g., $\boldsymbol{s} = (0, 1, 2, 3, 4)$ for the trait and $\boldsymbol{s}^{ERS} = (1, 0, 0, 0, 1)$ for ERS).

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

Model Comparison Using Empirical Data

In order to illustrate the different specifications, assumptions, and estimated parameters of the response style models, we analyzed a non-clinical standardization sample of a German Big Five inventory by Borkenau and Ostendorf (2008). In this sample, 11,724 respondents answered a Big Five questionnaire, wherein each scale consists of twelve items on a 5-point rating scale, hence 60 items in total.

As baseline models, we fit a PCM and a generalized PCM with discrimination parameters, both ignoring response styles, to the Big Five data. We chose the PCM and generalized PCM as a special case of the NRM with fixed scoring weights for the Big Five dimensions, as the (g)PCM reflects the ordinal structure of the

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

response categories, while a NRM with estimated scoring weights is rather suited to model responses to nominal categories (Thissen & Steinberg, 1986).

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

Response styles were modeled across all 60 Big Five items and with the same scoring for reversed and non-reversed items (see Table 1 and Wetzel & Carstensen, 2017, for a discussion on using the same, separate, or additional items for the response style dimension). All models were estimated using R (R Core Team, 2019) with the package *TAM* (Test Analysis Modules, Kiefer et al., 2017). Within *TAM*, we used the Marginal Maximum Likelihood method to estimate multidimensional IRT models with estimated or fixed scoring weights and discrimination parameters. For high dimensional models, *TAM* offers a quasi Monte-Carlo integration procedure (Pan & Thompson, 2007) that prevents the time intensive numeric integration.

Model Specification

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

$\boldsymbol{\theta}^{BigFive} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. The generalized PCM had 310 parameters (240 fixed item-threshold parameters, 60 discrimination parameters, 5 latent trait variances for the Big Five dimensions were fixed to 1, and 10 latent covariances between dimensions were estimated with $\boldsymbol{\theta}^{BigFive} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$.

Scoring weights to specify the random threshold model (Wang et al., 2006) are presented in Table 1. Here, 259 parameters where estimated (240 item-threshold parameters, 5 Big Five variances, 4 threshold variances, and 10 latent covariances between Big Five dimensions with $[\boldsymbol{\theta}^{BigFive}, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where covariances were fixed to 0 between Big Five dimensions and thresholds, as well as between random thresholds). The same scoring weights were used for the generalized random threshold model (adapted from Wang & Wu, 2011), in which we estimated 60 additional discrimination parameters for the Big Five dimensions, and 60 discrimination parameters for the random threshold dimensions. Hence, 370 parameters were estimated (240 item-threshold parameters, 120 discrimination parameters, 5 Big Five variances, and 4 threshold variances were fixed to 1 with $[\boldsymbol{\theta}^{BigFive}, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where, as before, 10 latent covariances between Big Five dimensions were estimated, while covariances were fixed to 0 between Big Five dimensions and thresholds, as well as between random thresholds)[2].

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

For the model with category preference parameters for response styles (Bolt et al., 2014), scoring weights for the Big Five and the category preference dimensions were fixed (see Table 1). This results in 285 estimated parameters (240 item-threshold parameters, 5 Big Five variances, 4 category preference variances and

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

36 latent covariances with $[\boldsymbol{\theta}^{BigFive}, \theta_1, ..., \theta_4] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$; the last category preference parameter can be derived from the others as across categories they sum to 0).

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

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

Model Fit

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

Overall, accounting for response styles clearly led to better model fit (all $\chi^2 \geq 30,182$, $p < .001$). Based on BIC, it appears that allowing for dependencies between person-specific threshold shifts (Bolt & Johnson, 2009; Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017) instead of accounting for response styles as random noise (Wang et al., 2006; Wang & Wu, 2011) additionally increased model fit. Similarly, allowing for latent covariances between traits and response style dimensions seems to be a sensible approach in this dataset as response style models using the $\Sigma = Diag$ restriction (Wang et al., 2006; Wang & Wu, 2011) had higher BIC values than response style models estimating latent covariances (Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017, with an exception of the multidimensional NRM by Bolt & Johnson, 2009, and Bolt et al., 2014). Finally, allowing that latent dimensions impact items differently by adding discrimination parameters to the model substantially increased model fit (Random Threshold Model vs. generalized Random Threshold Model: $\chi^2(111) = 9,778$, $p < .001$; multidimensional PCM vs. generalized multidimensional PCM: $\chi^2(173) = 14,623$, $p < .001$).

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

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

TABLE 2: Overview of Estimated Parameters and Model Fit Indices for the IRT Models

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

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

TABLE 3: Estimated Correlation Matrix in the Best Fitting Model (Generalized Multidimensional PCM by Falk & Cai, 2016; Variance of Latent Traits was Fixed to 1)

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

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

To conclude, in the Big Five standardization sample, a clear advantage of models specifying response styles a priori and therefore allowing for covariances between traits, between response styles, and between traits and response styles (models by Bolt et al., 2014; Falk & Cai, 2016; Wetzel & Carstensen, 2017) was found in the data. Besides the increased model fit in this dataset, IRT variants with a priori specified response styles have a straight-forward interpretation of response style dimensions and the relation between latent dimensions can be explored through the variance-covariance matrix $\mathbf{\Sigma}$. In addition, an advantage of models using item-specific discrimination parameters emerged. Such or similar comparisons between response style models can be useful tools to test specific assumptions on response styles. For example, one can examine whether response style dimensions impact items differently through comparing a multidimensional PCM and a generalized multidimensional PCM. Even though we found an advantage for such a model in the Big Five standardization sample, we would like to emphasize that this analysis is for illustrative purposes only, and that we had no a priori assumptions on response styles that we aimed to test with the aid of this analysis.

New Model Extensions

In the Big Five standardization sample, we found a superiority of models specifying response styles a priori and allowing for differential influences of the response style dimensions on single items. However, both model specifications come at a price. First, specifying response style a priori implies strong assumptions on response style specifications, namely symmetric threshold shifts for ERS and MRS around

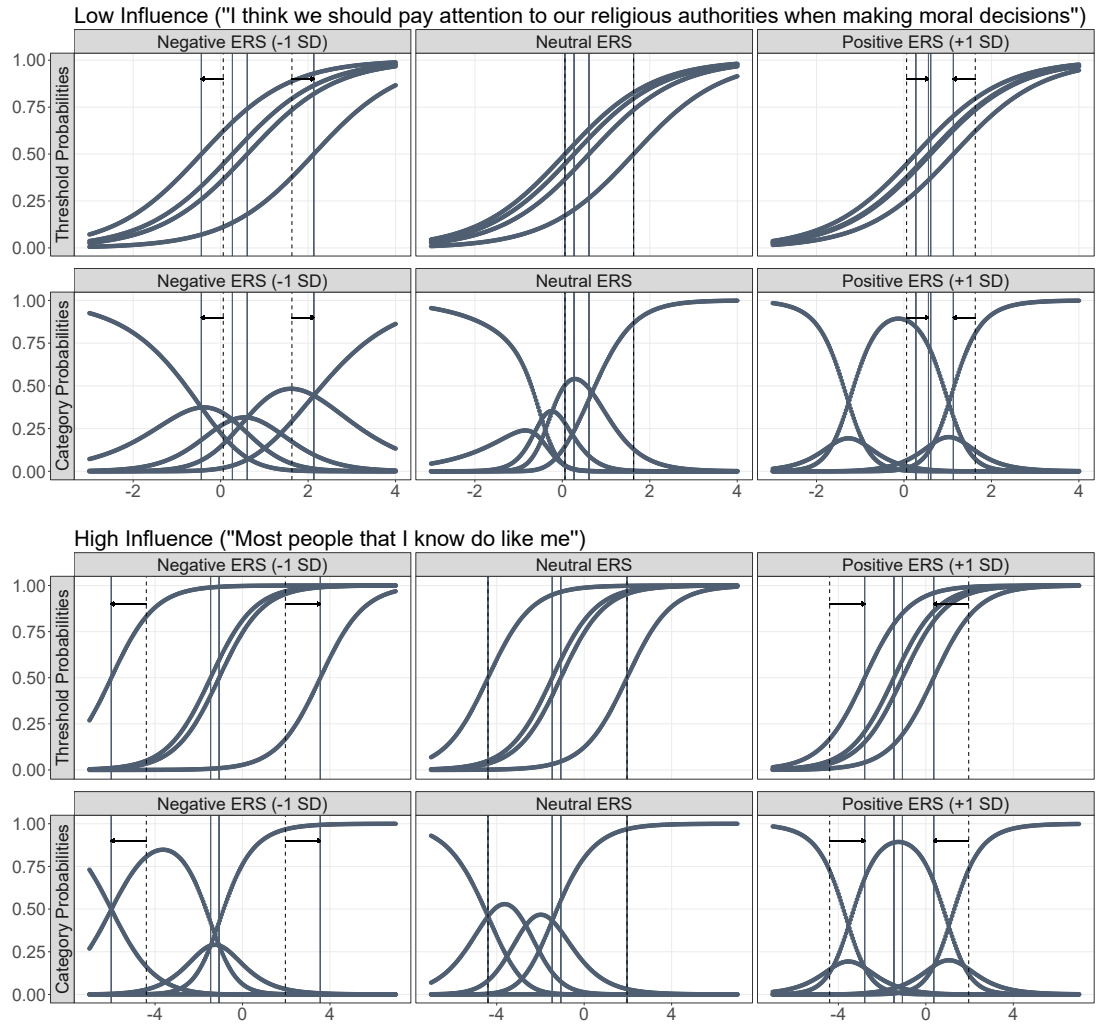

FIGURE 3: Illustration of the influence of low (upper panel) and high (lower panel) discriminability of the Extreme Response Style (ERS) dimension on threshold and category probabilities with model based item-threshold and discrimination parameters (Falk & Cai, 2016).

the item location ($\theta_n^{ERS} = -\delta_{n1} = \delta_{n4}$; $\theta_n^{MRS} = \delta_{n2} = -\delta_{n3}$ for an item with 4 thresholds). For ARS, scoring weights $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$ stand for a shift in the third threshold, while the threshold probability of the highest thresholds stays constant (see Figure 2 and Appendix B in Henninger & Meiser, 2019). Second, when including discrimination parameters for response style dimensions (Falk & Cai, 2016), the model becomes highly flexible through allowing the dimensions to have differential influences on the items. However, the number of estimated parameters increases tremendously, especially when the number of latent (response style) dimensions is large.

In this section, we propose two new model extensions that address these two

challenges: a model lifting the equality constraint on scoring weights (and therewith threshold shifts) in multidimensional PCMs for ERS, MRS, and ARS, and a model that reduces the number of estimated parameters by imposing equality constraints on discrimination parameters based on item attributes. Both models fill in gaps in the model structure. The model lifting equality constraints on threshold shifts in multidimensional PCMs is more flexible than fixing the scoring weights a priori (e.g., Wetzel & Carstensen, 2017), but uses a priori specifications of response styles in contrast to a multidimensional NRM (Bolt & Johnson, 2009). The second model that constrains discrimination parameters is more restrictive and parsimonious than the model by Falk and Cai (2016), but has a higher flexibility than a multidimensional PCM (e.g., Wetzel & Carstensen, 2017).

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

Lifting the Equality Constraint of Scoring Weights

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

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

The additional, estimated scoring weight parameter λ that is equal across participants and items reflects the assumption that effects of response styles on categories may not be the same for all categories, but proportional between categories within

items. For example for ERS, the extreme categories are not affected equally, but we can test whether the highest category is affected more strongly than the lowest category. When $\lambda^{ERS} > 1$, θ_n^{ERS} affects the highest agreement category more strongly than the lowest disagreement category and vice versa for $\lambda^{ERS} < 1$. $\lambda^{MRS} > 0$ implies that also the probability for intermediate categories increases for positive levels of θ_n^{MRS}. $\lambda^{ARS} > 1$ implies that θ_n^{ARS} influences the highest threshold and hence increases probability of choosing the highest category more strongly. Therewith, λ^{ARS} makes the assumption that ARS affects only certain threshold shifts testable (see the rightmost column in Figure 2 for shifts in threshold when $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, 1)$).

Modeling Discrimination Parameters Through Item Attributes

Response styles may have stronger or weaker influences on item responses depending on item attributes, such as item complexity or item position. To propose a more parsimonious model where the differential influence of the response style dimensions on items is captured by item-specific discrimination parameters α_{id}, we define item-specific discrimination parameters to be a function of item attributes. Such attributes can be contextual influences, such as the number of response options, item wording, ambiguity, complexity, negation, reversal, or position effects. For illustration, one can specify an explanatory IRT model in which the strength of response style effects is moderated by item complexity, negation and position:

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

The function f can be a linear parameter combination of item attributes; but also other kinds of function may apply. In the model we propose here, we allow heterogeneity of discrimination parameters α_{id} for items with different combination of item attributes, but restrict them to be equal within groups of items with the same combination of attribute levels (see below, and Table B1 in Appendix B). Hence, this model can be regarded as an explanatory IRT approach for discrimination parameters to investigate the impact of response style dimensions on specific item types (see De Boeck & Wilson, 2004, for more explanatory IRT approaches). Similar to Embretson (1999), the proposed model uses item attributes as predictors for discrimination parameters and follows a fixed-links approach (e.g., Schweizer, 2008; Zeller, Reiß, & Schweizer, 2017) such that parameters α_{id} are decomposed into elementary parameters. It hence tests the moderating role of item attributes on response style effects in a confirmatory way.

Fit of New Model Extensions to the Big Five Standardization Sample

We fit two exemplary specifications of the new modeling variants to the Big Five standardization sample. Of course, the approaches presented here serve as a guidance for applications and can be specified for any other latent dimension or adapted for other types of attributes or alternative explanatory approaches. For the fit of a model lifting the equality constraint on scoring weights, we used the response style dimension ARS as an example. Hence, we defined scoring weights for the ARS dimension as $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, \lambda^{ARS})$ to test whether and to which magnitude the ARS dimension affects the upper threshold. All other parameter were specified as in the model of Wetzel and Carstensen (2017) in the illustration section above. To fit a model with constrained discrimination parameters of response style dimensions ERS and MRS, we used three types of item attributes to define the restrictions: item negation, complexity and position (see Table B1 in Appendix B). Items received the value 1 when they were negated (e.g., contained "not", "not a", "never") and 0 otherwise; items were coded 1 on Complexity if the item content included more than one line of thought (i.e. double-bind items, e.g., "I am quite good at organizing my time for myself so that I can finish my affairs on time."). Please note that item responses in the 60 item version of the Big Five standardization sample used in for analyses herein were collected with a 240 item measure. We used the position of items from the 240 item instrument, so item received the value 1 when they occurred in the last half of the 240 item instrument, and 0 otherwise. The combination of the three dichotomous factors results in eight different combination of factor levels, therefore eight discrimination parameters α for each response style dimension ERS and MRS were estimated impacting items that met the combination of factor levels. All other parameters were specified as in the model of Falk and Cai (2016) in the illustration section including item-specific discrimination of content traits.

Table 4 extends the overview of estimated parameters and information criteria of the response style models (Henninger & Meiser, 2019) by the two exemplary modeling extensions. The model lifting the equality constraint of scoring weights from the ARS trait fits the data better than its restricted variant (Wetzel & Carstensen, 2017, $\chi^2(1) = 141$, $p < .001$). The scoring weights are $\boldsymbol{s}^{ARS} = (0, 0, 0, 1, \lambda^{ARS})$, with $\lambda^{ARS} = 1.36, SE < 0.01$. This indicates that for the ARS trait, not only the third threshold is shifted by θ_n^{ARS}, but also the upper threshold is shifted by $0.36 \cdot \theta_n^{ARS}$. Stated differently, this response style model variant shows that the threshold probability between the two agreement categories is affected by

TABLE 4: Overview of Estimated Parameters and Information Criteria for two new Exemplary Model Extensions

	Number of estimated						Model fit indices			
	parameters (total)	item-threshold parameters	latent variances	latent covariances	discrimination parameters		SGDDM	Log-Likelihood	BIC	LR-Test
Multidimensional PCM *(Lifting Equality Constraints from ARS Scoring Weights)*	277	$60 \times 4 = 240$	$5 + 3$ (0)	28	1	*1 ARS, (highest category)*	.046	-852,762	1,708,118	$\chi^2(1) = 141$ $p < .001$
Gen. Multidimensional PCM *(with Equality Constraints on Discrimination Parameters)*	345	$60 \times 4 = 240$	$5 + 3$ (8) *(all fixed to 1)*	28	77	*60 Big Five, 8 ERS/8 MRS, 1 ARS*	.042	-846,544	1,696,320	$\chi^2(69) = 12,577$ $p < .001$

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

TABLE 5: Estimated Discrimination Parameters for Each Factor Level Combination in the Generalized Multidimensional PCM with Equality Constraints on Discrimination Parameters

Negation	Complexity	Position	α^{ERS}	α^{MRS}
0	0	0	1.07 (0.01)	0.59 (0.01)
1	0	0	1.09 (0.01)	0.60 (0.01)
0	1	0	0.93 (0.01)	0.54 (0.01)
1	1	0	1.18 (0.02)	0.67 (0.02)
0	0	1	1.19 (0.01)	0.61 (0.01)
1	0	1	1.11 (0.02)	0.67 (0.02)
0	1	1	1.18 (0.01)	0.66 (0.01)
1	1	1	1.05 (0.03)	0.71 (0.03)

Note. ERS: Extreme Response Style, MRS: Mid Response Style, α: Discrimination Parameter; Standard Errors in parentheses.

the ARS trait, but to a lower degree than the threshold between the middle and the first agreement category.

In the new model variant improving constraints on discrimination parameters of the ERS and MRS latent traits, eight item discrimination parameters were estimated for each of the two response style dimensions (see Table 5). Hence, the new restrictions reduced the number of discrimination parameters from 60 to eight for each of the two response style dimensions (i.e. reducing 104 parameters in total). Unsurprisingly, the restricted model has a worse fit than its less restricted variant (Falk & Cai, 2016, $\chi^2(104) = 2,047$, $p < .001$), as we would not assume that the reduction in estimated parameters and model flexibility goes unnoticed. However, there is still a substantive increase in model fit compared to the multidimensional PCM with response styles and item-invariant discrimination parameters (Wetzel & Carstensen, 2017, $\chi^2(69) = 12,577$, $p < .001$) which speaks in favor of the utility of using information on item attributes for parameter estimation[3]. The results indicates that the impact of response styles on item responses is a function of item attributes and can be assessed with psychometric modeling approaches.

[3]As a competitor model, we fit an alternative approach where eight discrimination parameters were randomly assigned to the 60 Big Five items. In consequence, item characteristics could not have any systematic influence on discrimination of the latent traits. The competitor model fit the data worse ($\Delta BIC = -53$) than the model incorporating item characteristics further suggesting that variations of item attributes systematically affect the impact of response styles on item responses.

Discussion

A variety of IRT model extensions accounting for response styles can be subsumed under the superordinate framework of shifting thresholds (see Henninger & Meiser, 2019). Based on the framework, the models can be structured in three groups: models with unique individual profiles of response tendencies (Wang et al., 2006; Wang & Wu, 2011), models investigating the response style structure in the data exploratorily (Böckenholt & Meiser, 2017; Bolt & Johnson, 2009; Moors, 2003; Rost, 1991), and models specifying the structure of response styles a priori (Bolt et al., 2014; Bolt & Newton, 2011; Falk & Cai, 2016; Jin & Wang, 2014; Morren et al., 2011; Wetzel & Carstensen, 2017).

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

We illustrated model estimation and interpretation in a Big Five standardization sample (Borkenau & Ostendorf, 2008). Herein, we found a superiority of models that specified response styles a priori and therewith were able to estimate relations between latent dimensions, but also of models that allowed for a differential impact of latent dimensions on the items. Building on these results, we proposed two novel types of model extensions that add to the response style models (see Table 1 in Henninger & Meiser, 2019). The two exemplary models, first, showed to what magnitude ARS also affected the highest threshold and, second, illustrated how item attributes can inform discrimination parameters.

Alternative Approaches to Account for Response Styles

This research focused on IRT models for response styles that are multidimensional extensions of Divide-by-Total models. In the following, we give a brief overview on approaches based on the graded response model (Samejima, 1969), sequential (Tutz, 1997) or step models (Verhelst, Glas, & De Vries, 1997) and IRTree models (Böckenholt, 2012; De Boeck & Partchev, 2012) as well as design-based approaches to directly control for response styles during measurement.

Modeling approaches

Based on the graded response model, Rossi, Gilula, and Allenby (2001) introduced a proportional threshold model accounting for heterogeneity in response scale via

person-specific location as well as a scale parameter, dispersing or contracting the scale. Similarly, Johnson (2003) proposed a model wherein thresholds are symmetric around the midpoint of the scale and distances between threshold parameters vary between respondents. Both models focus on ERS. Based on Johnson (2003) and Rossi et al. (2001), Javaras and Ripley (2007) developed a multidimensional unfolding model, wherein thresholds can vary between groups unrestrictedly, or between individuals via shift and scaling parameters. De Jong, Steenkamp, Fox, and Baumgartner (2008) proposed an approach, where a general ERS dimension and trait dimensions are modeled simultaneously using the testlet model by Bradlow, Wainer, and Wang (1999).

Thissen-Roe and Thissen (2013) proposed a two-decision model based on the idea that respondents may take two steps to answer a Likert-type item. First, they decide whether they agree or disagree with the item, and second, how strongly they (dis)agree. This model is based on the GRM, includes item discrimination parameters and allows for only two dimensions (agreement and extreme response). More recent models specify covariates to disentangle trait and response style in a one-item, adjacent categories model (Tutz & Berger, 2016), or adapted the differential discrimination model (Ferrando, 2014) to ordinal responses (Lubbe & Schuster, 2017).

A prominent approach to modeling response styles are IRTree models (Böckenholt, 2012; De Boeck & Partchev, 2012) which represent responses to rating scale items as a sequence of multiple processes: whether the respondent gives a directional response or prefers the middle category (MRS), agrees or disagrees with the item (content), and gives an extreme or less extreme response (ERS; see also Khorramdel & von Davier, 2014, for a multi-scale extension; Meiser, Plieninger, & Henninger, 2019, for an extension to ordinal judgment processes; Plieninger & Heck, 2018, for an extension for ARS; Plieninger & Meiser, 2014, for a test of validity; Zettler, Lang, Hülsheger, & Hilbig,2016, for an application). Moreover, Jeon and De Boeck (2016) generalized the IRTree approach to accommodate different IRT models in each process, introduced a bifactor model for multiple dimensions and included covariates in the model.

Design-based approaches

In case that researchers can influence the measurement situations, adapting measurement methods may be a promising tool to control response style impact. For example, situational factors such as respondents' motivation or cognitive load (Cabooter, 2010), or features of the questionnaire format such as the number of

categories, response option labels, reverse-coded or negated items (Weijters et al., 2010b) may reduce response biases. In the multidimensional forced-choice format, for instance, respondents rank groups (e.g., triplets) of items depending on how well they describe their behavior. Data from this format is ipsative by nature, which can be resolved by using a Thurstonian IRT model (Brown & Maydeu-Olivares, 2013). Alternatively, McKeown and Thomas (1988) proposed the Q-Methodology wherein respondents are asked to sort items into categories with prespecified assignment rates per category. The sorting result reflects a normal distribution with the middle category containing most items, flattening towards the tails. Similarly, Böckenholt (2017) used a method proposed by Thurstone (1928) asking respondents to sort items into categories using a drag-and-drop procedure. Current research focuses on the power to reduce response style effects by these and other response formats (see for example Plieninger, Henninger, & Meiser, 2019, for an experimental investigation of the drag-and-drop format).

Directions for Future Research

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

The advancement of existing models is a further route for future research. For instance, the random threshold model by Wang and colleagues (Wang et al., 2006; Wang & Wu, 2011) is a promising candidate for modeling response styles as it allows researchers to model heterogeneity towards any response category

with little a priori assumptions. This is particularly important when comparing different subgroups with unknown response styles, as might, for example, be the case in cross-cultural research. However, as demonstrated in the application, the model is likely to be violated in empirical data due to the independence restriction on the variance-covariance matrix. Furthermore, it is not possible to interpret person-specific threshold effects in terms of ERS or MRS, because then response styles induce a non-diagonal variance-covariance matrix of person effects. Therefore, more flexibility in the random threshold model concerning its identification constraints is desirable, allowing to estimate the variance-covariance matrix.

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

Besides advancing estimation and modeling approaches, a substantive analysis of discrimination parameters of response style dimensions may help to identify sources of biases and problematic items in test construction. Discrimination parameters indicate item-specific differences in the strength of response style effects on item responses and hence indicate which items are more strongly affected by response style traits (see Falk & Cai, 2016). Testing hypotheses about moderating item attributes, such as ambiguity, item position, or complexity will provide valuable information to improve item generation and selection in test construction. Such specific hypothesis-based tests can lead to a reduction of the systematic impact of response styles on category choices and therewith biases in social science measurement situations (Podsakoff, MacKenzie, & Podsakoff, 2012).

Conclusion

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

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

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

References

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

Asparouhov, T., & Muthén, B. (2012). General random effect latent variable modeling: Random subjects, items, contexts, and parameters, Paper presented at the third UK Mplus Users' Meeting, London, UK. Retrieved from http://www.statmodel.com/download/NCME12.pdf

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559. doi:10.1007/s11336-008-9092-x

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

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115. doi:10.1509/jmkr.45.1.104

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

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

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

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

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

Henninger, M., & Meiser, T. (2019). Different approaches to modeling response styles in Divide-by-Total IRT models (Part I): A model integration. *Invited Revision Submitted to Psychological Methods*.

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

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

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

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

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

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

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

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance

for dimensionality assessment for multidimensional models. *Journal of Educational Measurement*, *52*, 144–158. doi:10.1111/jedm.12070

Lubbe, D., & Schuster, C. (2017). The graded response differential discrimination model accounting for extreme response style. *Multivariate Behavioral Research*, 1–14. doi:10.1080/00273171.2017.1350561

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thissen-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

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

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

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

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

Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the Partial Credit Model. *Applied Psychological Measurement*. doi:10.1177/0146621617748322

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

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

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

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

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

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

Appendix A

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

Trait 1

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

Trait 2

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

ERS

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

MRS

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

Appendix B

Coding of Item Characteristics for the 60 Big Five Items

TABLE B1: Coding of Item Negation, Complexity, and Position for the Generalized Multidimensional PCM with Equality Constraints on Discrimination Parameters

Item 1 - 30					Item 31 - 60				
Item	Negat.	Complex.	Posit.	Param.	Item	Negat.	Complex.	Posit.	Param.
N 1	1	0	0	α_1	N 31	0	0	0	α_2
E 2	0	0	0	α_2	E 32	0	0	0	α_2
O 3	1	0	0	α_1	O 33	0	1	1	α_6
A 4	0	1	0	α_3	A 34	0	0	1	α_4
C 5	0	0	0	α_2	C 35	0	0	0	α_2
N 6	0	0	1	α_4	N 36	0	1	0	α_3
E 7	0	0	1	α_4	E 37	0	0	1	α_4
O 8	0	0	0	α_2	O 38	0	1	0	α_3
A 9	0	0	1	α_4	A 39	0	0	0	α_2
C 10	0	1	0	α_3	C 40	0	1	1	α_6
N 11	0	1	0	α_3	N 41	0	1	1	α_6
E 12	1	0	1	α_5	E 42	1	0	0	α_1
O 13	0	1	0	α_3	O 43	0	1	1	α_6
A 14	0	0	0	α_2	A 44	0	0	0	α_2
C 15	1	0	0	α_1	C 45	1	1	0	α_7
N 16	0	0	0	α_2	N 46	0	0	0	α_2
E 17	0	0	1	α_4	E 47	0	0	1	α_4
O 18	0	1	0	α_3	O 48	0	1	1	α_6
A 19	0	1	0	α_3	A 49	0	0	0	α_2
C 20	0	0	0	α_2	C 50	0	1	0	α_3
N 21	0	0	0	α_2	N 51	0	1	0	α_3
E 22	0	0	1	α_4	E 52	0	0	1	α_4
O 23	1	0	1	α_5	O 53	0	0	1	α_4
A 24	0	0	0	α_2	A 54	1	1	0	α_7
C 25	0	1	0	α_3	C 55	1	1	1	α_8
N 26	0	0	0	α_2	N 56	0	1	0	α_3
E 27	0	0	0	α_2	E 57	0	1	1	α_6
O 28	0	0	0	α_2	O 58	0	0	0	α_2
A 29	0	1	0	α_3	A 59	0	1	0	α_3
C 30	0	1	0	α_3	C 60	0	0	1	α_4

Note. Negat.: Negation; Complex.: Complexity, Posit.: Position (based on the 240 item measure), Param.: Parameter; N: Neuroticism; E: Extraversion; O: Openness; A: Agreeableness; C: Conscientiousness.

A Novel Varying Threshold IRT Approach to Accounting for Response Styles

Mirka Henninger

University of Mannheim

Abstract

IRT models with varying thresholds are essential tools to account for unknown types of response styles in rating data. However, in order to separate content traits to be measured and response tendencies, specific constraints have to be imposed on varying thresholds and their interrelations. A sum-to-zero constraint for varying threshold models is proposed that allows us to flexibly account for response tendencies and to model covariations between varying thresholds that are commonly found in empirical data. The model's ability to capture different kinds of response tendencies is shown in a simulation study. An illustrative multi-country analysis demonstrates that the new model is well suited to account for extreme and mid response styles, but also for unknown, previously unmodeled, response tendencies.

Rating scales are in widespread use to measure personality, attitudes, and beliefs in psychological and educational assessment settings. They are common psychological measurement tools to assess the Big Five personality factors (e.g., Costa & McCrae, 2008) or background information in educational measurement studies such as the Program for International Student Assessment (PISA) or the Programme for the International Assessment of Adult Competencies (PIAAC).

When persons give responses to rating scale items, they do not only differ in terms of the content trait to be measured, but also with respect to the way they use the rating scale (Baumgartner & Steenkamp, 2001). The so-called response styles can be regarded as latent traits that describe heterogeneity in response scale usage and predict respondents' tendencies towards choosing certain kinds of categories. Types of response styles identified in the literature are *extreme response style* (ERS, a tendency for the highest and lowest categories), *mid response style* (MRS, a tendency towards the middle category) and *acquiescence response style* (ARS, a tendency to agree with the item; see Van Vaerenbergh & Thomas, 2013).

Response styles seem to be omnipresent in rating scale data. Different applications of mixture distribution models have shown that approximately one third of respondents give more extreme responses to rating scale items, whereas two thirds use the moderate response options more often (Eid & Rauber, 2000; Meiser & Machunsky, 2008). Similarly, in models with continuous response style dimensions, ERS has been found to possess substantial variance (see Böckenholt & Meiser, 2017; Wetzel, Böhnke, & Brown, 2016). Furthermore, response styles seem to be largely independent of item content (see Van Vaerenbergh & Thomas, 2013; Weijters, Cabooter, & Schillewaert, 2010; Wetzel, Carstensen, & Böhnke, 2013) and to be persistent over time (Weijters, Geuens, & Schillewaert, 2010; Wetzel, Böhnke, & Brown, 2016). Hence, they can be considered a systematic source of error in measurement. As a consequence, response styles can bias conclusions drawn from measurement in terms of measurement precision (Bolt, Lu, & Kim, 2014; Wetzel & Carstensen, 2017), relations between variables (Böckenholt & Meiser, 2017), or cross-group comparisons, for example in cross-cultural research (Bolt et al., 2014; G. W. Cheung & Rensvold, 2000; Harzing, 2006; Morren, Gelissen, & Vermunt, 2012).

To reduce distorting influences of response styles on measurement, a variety of methods for questionnaire design and psychometric modeling have been proposed. Researchers have examined the potential of the number of categories, labels, reverse-coded, or negated items (Weijters, Geuens, & Schillewaert, 2010) or alternative response formats (Böckenholt, 2017; Brown & Maydeu-Olivares, 2013) to

reduce response styles during measurement. To account for confounding influences of response styles in a given dataset, response style parameters are added to Item Response Theory (IRT) models. Different variants of psychometric approaches account for unknown response tendencies in rating data (Wang, Wilson, & Shih, 2006; Wang & Wu, 2011), allow us to explore response styles (Bolt & Johnson, 2009; Rost, 1991), or to investigate pre-specified response styles in terms of their relation to other variables or their impact on single item responses (Böckenholt, 2012; Falk & Cai, 2016; Wetzel & Carstensen, 2017).

The focus of this article lies on psychometric modeling approaches accounting for response styles and their explicit or implicit assumptions on the type of response styles. When little is known about the types of response styles in the data, or when the types of response tendencies differ between subgroups of respondents, strong a priori assumptions on the type of response styles are likely to be violated. Therefore, in such cross-group settings, flexible modeling approaches are needed to correct for confounding effects of response biases.

The goal of this article is to propose a novel modeling approach that is based on varying thresholds to account for response styles. The approach allows for the modeling of response styles that are commonly present in rating data, such as ERS or MRS, but at the same time retains the flexibility to accommodating unknown response tendencies. In the remainder of this article, I will propose the characteristics of such a model and distinguish it from existing approaches. Finally, I will examine the ability of the new approach to estimate trait and response style parameters in a simulation study and illustrate the approach with a multi-country analysis using rating scale measures of the Big Five personality factors.

Response Styles in IRT Approaches

In adjacent category IRT models like the Partial Credit Model (PCM; Masters, 1982), item responses can be modeled through threshold and category probabilities. The threshold probability is defined as the conditional probability of choosing category k when the response is either in category k or $k-1$. It is given by

$$p(X = k|X \in \{k-1, k\}, \theta, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{\exp\left(\theta_n - (\beta_i + \tau_{ik})\right)}{1 + \exp\left(\theta_n - (\beta_i + \tau_{ik})\right)} \tag{1}$$

and is as a function of the trait parameter θ_n for person n, the item parameter β_i and the threshold parameter τ_{ik} for item i and category k. The category probability

formula of a PCM for $K+1$ categories with $k \in \{0, ..., K\}$ is given by

$$p(X = k|\theta, \boldsymbol{\beta}, \boldsymbol{\tau}) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k}(\beta_i + \tau_{ik'})\right)}{\sum_{j=0}^{K}\exp\left(s_j\theta_n - \sum_{k'=0}^{j} -(\beta_i + \tau_{ik'})\right)} \tag{2}$$

with $s_0\theta_n - (\beta_i + \tau_{i0}) \equiv 0$ and $\sum_{k=1}^{K}\tau_{ik} = 0$. The category or scoring weights s_k describe the relation between trait and category and are usually fixed to $\boldsymbol{s} = (0, 1, ..., K)$ in a PCM.

When there are response tendencies in the data, not all covariances between rating responses are captured by the model parameters. Then the remaining covariances due to response styles must be accounted for by additional model parameters, for example through person-specific shifts in threshold parameters δ_{nk} (e.g., Adams, Bolt, Deng, Smith, & Baker, 2019; Bolt et al., 2014; Wang et al., 2006):

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

or

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

Herein, δ_{nk} indicates a shift of person n for threshold k increasing or decreasing the probability that a certain category is chosen. The latent traits follow a multivariate normal distribution with $[\theta, \delta_1, ..., \delta_K] \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$. Such person-specific threshold shifts now allow us to capture response tendencies in the data. For example, through person-specific threshold shifts, an IRT model can allow for preferences for the extreme categories ERS: the outer thresholds are shifted towards the item location, widening the area over which the extreme categories have the modal category probability (see column 1 in Figure 1). Similarly, a preference for the middle category can be incorporated through outward threshold shifts of the inner thresholds (see column 2 in Figure 1). But, also more unsystematic and

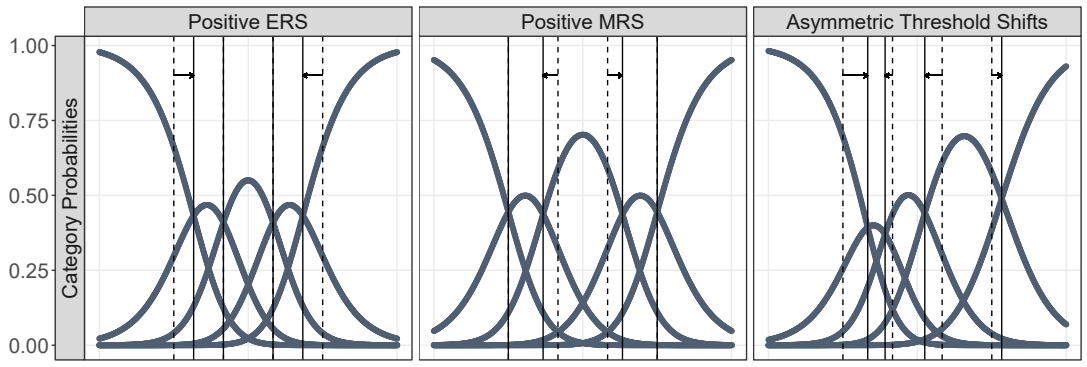

FIGURE 1: Category probability curves for an item with 5 response categories and 4 thresholds (vertical bars). From left to right: accounting for ERS through inward shifts of outer thresholds, accounting for MRS through outward shifts of inner thresholds, accounting for more unsystematic response tendencies through asymmetric threshold shifts.

asymmetric response tendencies can be captured by threshold shifts (see rightmost column in Figure 1).

Of course, restrictions must be imposed on either δ_{nk} or $\boldsymbol{\Sigma}$ in order to identify the IRT model with person-specific threshold shifts in estimation. This way, redundancies between traits and varying thresholds are avoided (e.g., a case where all thresholds are shifted towards one direction). Different restrictions have been implemented for different models in the response style literature. For example, models including ERS and MRS usually assume a perfect negative correlation of the outer thresholds (see column 1 and 2 in Figure 1; e.g., Falk & Cai, 2016; Wetzel & Carstensen, 2017) to separate content trait from response style dimensions. In contrast, Wang et al. (2006) proposed an approach with independent varying thresholds between respondents by using a restriction on $\boldsymbol{\Sigma} = Diag$, hence uncorrelated person-specific threshold shifts. In consequence, correlated thresholds—as is the case for ERS or MRS—constitute a violation of the independence assumption in such models.

Sum-to-zero Constraint on Threshold Variances Across Items

Hence, incorporating response styles into IRT approaches requires specific a priori assumptions with respect to response styles and covariations between latent traits. In this article, I propose a model that refrains from imposing strong restrictions on relations between varying thresholds as is the case in multidimensional PCMs (such

as perfect negative correlations between shifts of outer thresholds, e.g., Falk & Cai, 2016; Wetzel & Carstensen, 2017) and the random threshold model (independent threshold shifts, e.g., Wang et al., 2006).

The main assumption of the new modeling approach is that person-specific thresholds δ_{nk} sum to zero across thresholds within respondents:

$$\sum_{k=1}^{K} \delta_{nk} = 0 \; \forall \; n. \tag{5}$$

In this model, response styles are incorporated as person-specific shifts in threshold parameters, and thus do not have to be defined a priori. At the same time, the model implicitly incorporates model-implied dependencies between varying thresholds that are typically found in empirical data (e.g., Bolt & Johnson, 2009; Bolt et al., 2014; Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel & Carstensen, 2017). Incorporating dependencies between varying thresholds while keeping the flexibility to model unknown response styles is particularly important for educational measurement settings where researchers may not have a priori knowledge on the type of response styles in the data. For example, in cross-cultural settings, researchers may like to test whether different countries possess different types of response tendencies and whether specifying certain response styles a priori may be too restrictive.

Consequences of using a sum-to-zero constraint on varying thresholds

The new sum-to-zero constraint on varying thresholds has several characteristics. I will briefly outline these characteristics and return to each in the following section. First, in the model each respondent has a unique profile of threshold shifts that may vary in their quality as well as magnitude. This makes the model very flexible with regards to modeling unknown response tendencies in the data. Second, threshold shifts have no impact on item difficulty. Across all categories, δ_{nk} do not add or subtract to the linear parameter combination (Equation 3). Therefore, the location of the respondent on the latent continuum is set by the content trait. Third, as varying thresholds sum to zero within respondents, the constraint implicitly incorporates dependencies between varying thresholds that are often found in empirical data.

(1) Individual respondents' response style profiles

The sum-to-zero constraint allows us to make intraindividual statements about the relative strengths of threshold shifts within respondents by interpreting the ordering, direction, and magnitude of threshold shifts. Hence, we can interpret the individual profile of each respondent indicating which thresholds are shifted to which direction to what amount (Cornwell & Dunlap, 1994; Rost, 2004). These threshold shifts reflect the respondents' perception of the rating scale and we can interpret the shifts in terms of their cognitive representation of category width. The variance of sum-to-zero threshold shifts between persons can be interpreted as a variance of relative measures. It indicates the heterogenity between respondents in the relative magnitude of threshold shifts.

(2) Modeling of response styles

As the sum-to-zero constraint has no impact on item difficulty, it does not allow for the modeling of response styles that make items easier or more difficult. For example, it is not possible to account for acquiescence as it requires that agreement categories become more probable. Similarly, completely independent thresholds shifts would violate the sum-to-zero assumption as the constraint reduces the number of independent shifts by one and therewith enforces dependencies between varying thresholds.

(3) Ipsatized threshold shifts

Using a sum-to-zero constraint has implications on covariances between varying thresholds. For rating scale responses to $K+1$ response categories with K varying thresholds, only $K-1$ threshold shifts are free, while the K^{th} threshold shift is determined by the other $K-1$ threshold shifts. Thus, the constraint leads to *ipsatized* varying thresholds. At the same time, the constraint avoids redundancy of $K-1$ varying thresholds and the content trait(s) which allows us to estimate variances and covariances between content traits and $K-1$ varying thresholds. As $\delta_K = -\sum_{k=1}^{K-1} \delta_{nk}$, the variance and covariances with respect to the K^{th} threshold can be calculated from the covariance matrix $\mathbf{\Sigma}$ using a conversion of covariances:

$$\begin{aligned} \mathrm{Var}(\delta_K) &= \sum_{k=1}^{K-1} \mathrm{Var}(\delta_k) + 2 \sum_{k=1}^{K-2} \sum_{k'=k+1}^{K-1} \mathrm{Cov}(\delta_k, \delta_{k'}) \\ \mathrm{Cov}(\delta_k, \delta_K) &= -\sum_{k'=1}^{K-1} \mathrm{Cov}(\delta_k, \delta_{k'}). \end{aligned} \tag{6}$$

Hence, the full variance-covariance matrix of K varying thresholds is rank deficient, as the K^{th} row or column is a linear combination of the other rows or columns. Therefore, the expected correlations between varying thresholds is not zero, but negative and amounts to

$$\hat{r}_{k,k'} = \frac{-1}{K-1} \tag{7}$$

where K is the number of thresholds (see Chan, 2003; Clemans, 1966; Dunlap & Cornwell, 1994; Radcliffe, 1963). For example, if the K thresholds are uncorrelated or correlated to an equal amount, for $K = 4$ the expected correlation amounts to $\hat{r}_{k,k'} = -1/3$, and the expected correlations become smaller, the higher the number of thresholds. Furthermore, in the $K \times K$ covariance matrices at least one of the $K-1$ covariances between varying thresholds must be negative as the sums of the rows or columns in the covariance matrix are equal to zero (Chan, 2003; M. W.-L. Cheung, 2004).

A covariance matrix based on ipsatized data cannot be corrected to its non-ipsatized counterpart without the assumption that non-ipsatized variables are uncorrelated and homoscedastic (Chan, 2003). Besides, it is not possible to apply factor analytic approaches in order to assess dominant factors among varying thresholds to the rank-deficient $K \times K$ covariance matrix (M. W.-L. Cheung, 2004). Therefore, we cannot interpret the absolute height of correlations among varying thresholds. However, correlations can be interpreted in terms of the rank order of correlations (Rost, 2004), indicating which correlations are smaller or larger than others.

Distinction from other adjacent category IRT models for response styles

The model using a sum-to-zero constraint adds to the psychometric modeling approaches accounting for response styles. It differs from existing approaches by model-implied characteristics of person-specific threshold shifts δ_{nk} and their implications on the covariance matrix $\boldsymbol{\Sigma}$.

Compared to mixture distribution models (Rost, 1991) with one set of threshold shifts per latent class of respondents, threshold shifts in the model with sum-to-zero constraint are person-specific. In the threshold dispersion model by Jin and Wang (2014), thresholds are pushed apart or pulled together by the response style trait. However, the quality of shifted thresholds is the same for all respondents and reflects a mixture of ERS and MRS, while only the magnitude of response style is person-specific. In contrast, in the model proposed here, each respondent has individual threshold shifts that may vary in their direction as well as strength. Similarly, in multidimensional NRMs (e.g., Bolt & Johnson, 2009), person-specific threshold shifts are condensed into one or few response style dimensions. In consequence, the impact on single thresholds is defined through estimated scoring weights that are equal for all respondents, where only the magnitude of response style influence varies between respondents through the response style trait dimension(s). Moreover, the sum-to-zero constraint on K varying thresholds differs substantially from a sum-to-zero constraint imposed on the $K+1$ category preference parameters that was proposed by Bolt et al. (2014). While in the model presented here the constraint on varying thresholds δ_{nk} fixes the location of the respondent on the latent continuum of the target trait, this is not the case for category preference parameters in the model by Bolt et al. (2014). As an example, acquiescence can be incorporated in the model by Bolt et al. (2014) through person-specific category preferences $\boldsymbol{\theta}_n = (-1, -1, -1, 1.5, 1.5)$ for a 5-category item. Here, the probability for the agreement categories increases, thus the response style effect adds to the location of respondent n on the latent continuum. This decreases item difficulty for respondents with positive ARS trait levels which would not possible when using a sum-to-zero constraint on varying thresholds.

In particular, the novel model with a sum-to-zero constraint on varying thresholds closes a gap between two modeling approaches for response tendencies: a multidimensional PCM for ERS and MRS (Bolt & Johnson, 2009; Falk & Cai, 2016; Wetzel & Carstensen, 2017) and a random threshold model (Wang et al., 2006). In a multidimensional PCM, person-specific threshold shifts δ_{nk} are restricted a priori (Bolt & Newton, 2011; Falk & Cai, 2016; Wetzel & Carstensen, 2017) to include specific response style dimensions such as ERS into the IRT model. For example, when ERS is accounted for, the first and last threshold are shifted to the item location by the same amount ($\boldsymbol{\delta}_n = (-1 \cdot \theta_n^{ERS}, 0, 0, 1 \cdot \theta_n^{ERS})$; see column 1 in Figure 1). Hence, a restriction on threshold correlations is imposed into the model so that $\rho(\delta_1, \delta_4) = -1$ for ERS, and $\rho(\delta_2, \delta_3) = -1$ for MRS in a

5-category item. In the random threshold model proposed by Wang et al. (2006) the variance-covariance matrix $\boldsymbol{\Sigma}$ is restricted to a diagonal matrix to separate content traits from response style effects. In this model, no common structure of response tendencies across respondents can be modeled.

Estimation in Standard Software

The varying threshold model using a sum-to-zero constraint can be written as a multidimensional PCM with varying thresholds as $K - 1$ additional trait dimensions. In consequence, the model can be estimated with standard IRT software programs using Marginal Maximum Likelihood Estimation. The model reformulation as a multidimensional PCM is provided in Appendix A.

Present Research

In the remainder of the article, a simulation study demonstrates the ability of the new varying threshold model with sum-to-zero constraint to estimate item-threshold, content trait and response style parameters. Furthermore, data of a Big Five personality questionnaire from the Open Source Psychometrics Project (2019) is used to illustrate the application of the novel model with sum-to-zero constraint on varying thresholds. In a multi-country analysis, I highlight the differences between the new model implementing the sum-to-zero constraint on varying thresholds, a multidimensional PCM accounting for ERS and MRS, and a random threshold model in terms of response style specification and estimation. The analysis demonstrates that besides capturing systematic, known response tendencies, the novel model allows us to test whether unknown response tendencies are present in data originating from psychological assessments.

Simulation Study

The model using a sum-to-zero constraint on varying thresholds can be estimated as a multidimensional PCM (see Appendix A). As the ability of the model to estimate model parameters in different data scenarios is unknown, a simulation study was conducted to ensure that content trait and response style parameters are estimated well under different data structures such as different numbers of content trait dimensions and different numbers of reversed-coded items.

Setup for Data Generation and Model Fit

Item responses were simulated according to Equation 3 and 5 for $N = 500$ respondents answering $I = 12$ items per content trait dimension to a 5-point rating scale. The number of content trait dimensions ($N_{ContentTrait} \in (1, 2, 3)$), and the number of reversed-coded items per content dimension ($N_{Reversed-Coded} \in (0, 2, 4, 6)$) were varied, resulting in 0, $^1/_6$, $^1/_3$, or $^1/_2$ of the items per content dimension being reversed-coded. Through this setup, one can examine whether several content dimensions are needed to validly measure varying thresholds (see Wetzel & Carstensen, 2017) and determine the necessary number of reversed-coded items (see Plieninger, 2017).

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 $U(-2.5, 2.5)$ and centered. Person parameters for content trait(s) and varying thresholds were drawn from a multivariate normal distribution with $MVN(\mathbf{0}, \mathbf{\Sigma})$ and varying thresholds were centered afterwards to reflect the sum-to-zero constraint. The variances of content trait(s) were set to $\boldsymbol{\sigma}_\theta^2 = 1$, the variances of thresholds to $\boldsymbol{\sigma}_\delta^2 = (1, 0.5, 0.5, 1)$ to reflect that usually variances of outer thresholds are larger (e.g., Wang et al., 2006). The off-diagonal elements of $\mathbf{\Sigma}$ were drawn from a Wishart distribution with $df = 10$ and scale matrix $\mathbf{\Sigma}^*$ (here for 3 content dimensions):

$$\mathbf{\Sigma}^* = \left(\begin{array}{ccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & -.5 \\ 0 & 0 & 0 & 0 & 1 & -.5 & 0 \\ 0 & 0 & 0 & 0 & -.5 & 1 & 0 \\ 0 & 0 & 0 & -.5 & 0 & 0 & 1 \end{array}\right)$$

The scale matrix mirrors the strong negative correlations between the outer and between the inner thresholds, respectively, reflecting ERS and MRS (with 90% of correlations in the interval $[-0.86, -0.67]$ for ERS and $[-0.74, -0.45]$ for MRS; due to the difference in variance), but small correlations between traits, and traits and varying thresholds (90% of correlations in the interval $[-0.25, 0.25]$). The outer and the inner thresholds had an expected level of negative correlations (with 90% of correlations in the interval $[-0.66, 0.04]$; see also Plieninger & Heck, 2018).

In the simulation study the novel model with varying thresholds that summed to zero was fit to the generated data. The model was specified as a multidimensional PCM with scoring weights according to Table A2 in Appendix A with

varying thresholds affecting item response to an equal amount across all content dimensions (see Wetzel et al., 2013, for a discussion on consistency of response styles across content trait dimensions). The *R* package *TAM* (Kiefer, Robitzsch, & Wu, 2017) with Marginal Maximum Likelihood Estimation and a quasi Monte-Carlo integration procedure was used for model fit[1]. $R = 1000$ replications were realized for each condition ($N_{ContentTrait} \in (1, 2, 3)$, $N_{Reversed-Coded} \in (0, 2, 4, 6)$). Estimation of model parameters were evaluated 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.

Results

Panel A in Figure 2 shows histograms of the correlation between true and estimated content trait parameters for each condition and replication, Panel B shows the bias of estimated content trait parameters. I used a linear model to predict Fisher z-standardized correlation and bias averaged across content dimensions by the number of reversed-coded items and the number of content traits. The level of significance was set to $\alpha = .001$.

The number of reversed-coded items increased the correlation between true and estimated parameters ($b = 0.01, t = 46.49, p < .001$), but effects were minor in size. The number of content traits did not have an effect ($b = -0.00, t = -2.23, p = .026$). On average, the correlation between true and estimated parameters was $M_{Correlation} = 0.93$ similar to comparable simulation studies (e.g., Plieninger & Heck, 2018; Wetzel, Böhnke, & Rose, 2016). There were no effects of the predictors on bias (all $|b| < 0.01$, all $|t| < 1.96$, all $p \geq .050$), with an average bias $|M_{Bias}| < 0.01$.

Panel C in Figure 2 shows histograms of the correlation between true and estimated varying threshold parameters for each condition and replication, Panel D shows bias of estimated varying threshold parameters. Overall, parameter recovery was better for content traits than for varying thresholds. The outer thresholds are recovered better than the inner thresholds which is due to their higher variance. Fisher z-standardized correlations between each of the true and estimated varying thresholds were used as the dependent variables in a multivariate linear model. The model showed that the number of reversed-coded items decreased the correlation between true and estimated parameters (all $b \leq -0.06$, all $t \leq -144.10$, all $p <$

[1]Furthermore packages dplyr (Wickham, François, Henry, & Müller, 2018), ggplot2 (Wickham, 2016), gridExtra (Auguie, 2017), here (Müller, 2017), MASS (Venables & Ripley, 2002), MBESS (Kelley, 2018), truncnorm (Mersmann, Trautmann, Steuer, & Bornkamp, 2018) were used for data generation, data management, and plotting.

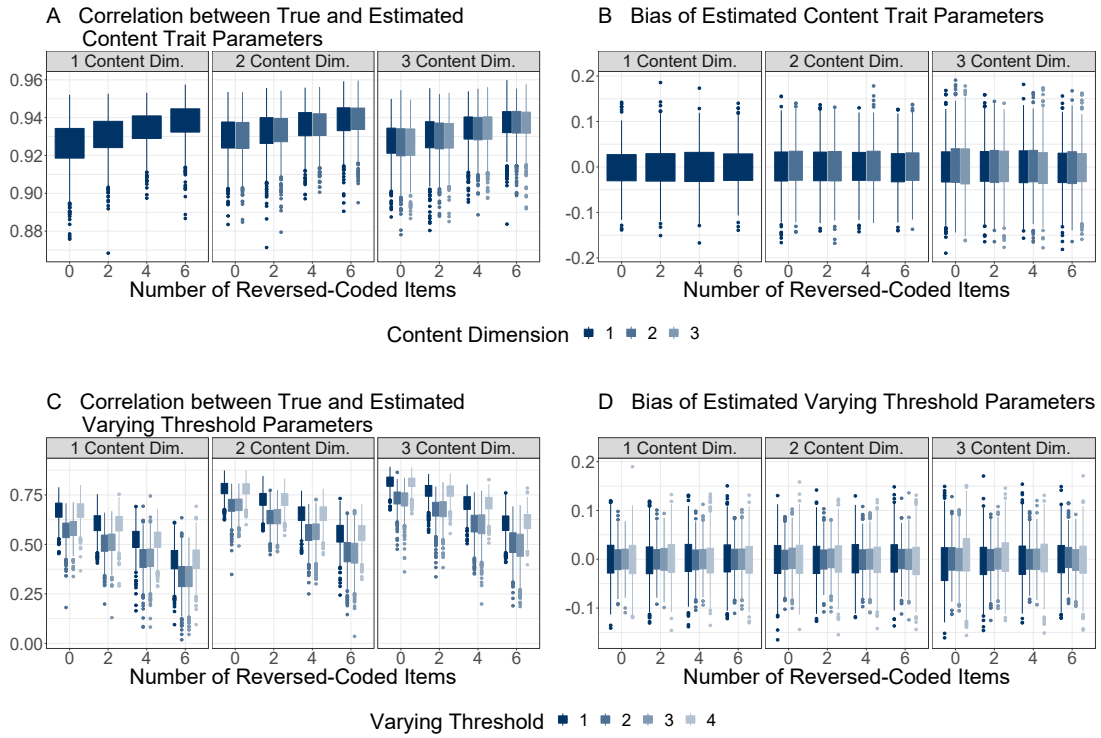

FIGURE 2: (A) Correlation between true and estimated content trait parameters; (B) Bias of estimated content trait parameters; (C) Correlation between true and estimated varying threshold parameters; (D) Bias of estimated varying threshold parameters as a function of the number of reversed-coded items and the number of content trait dimensions.

.001), while the number of content traits increased the correlation (all $b \geq 0.12$, all $t \geq 104.20$, all $p < .001$). There were minor effects of the number of reversed coded items and of content trait dimensions on bias for the outer thresholds ($-.01 < b < 0.01$, all $|t| < 3.88$, all $p < .001$), but no effects on the inner thresholds ($-.01 < b < 0.01$, all $|t| < 2.28$, all $p \geq .023$).

In conclusion, content trait parameters were recovered well without notable bias. While the number of reversed-coded items marginally increased parameter estimation, the number of content traits did not have any effect on estimation of content traits. Thus, the model using a sum-to-zero constraint can validly estimate content trait parameters even when only few or no reversed-coded items are present in the data or when they are modeled based on one content dimension. Similarly, there is no notable bias in the estimation of varying thresholds.

Model Illustration in a Multi-Country Setting

The new model using a sum-to-zero constraint allows us to assess confounding influences of response tendencies and test whether response tendencies diverge from common response styles such as ERS or MRS. In this illustrative analysis, the model is used in a multi-country setting to test whether country-specific differences with respect to response tendencies are present in data.

The dataset used in this analysis originates from the Open Source Psychometrics Project (2019) which offers anonymous data from various psychological constructs such as personality and attitude measures for psychometric research. The dataset contains responses to a Big Five questionnaire from the International Personality Item Pool (Goldberg, 1992) collected in various countries. Each of the five scales (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness) contained ten items with at least two reversed-coded items per dimension. Responses to the scales were given on a 5-category rating scale. Paying careful attention to data validity, only respondents with complete answers and who chose at least three out of the five response categories across all items were included into the analyses. In addition, only respondents whose native language was English were included to avoid confounds of the test language. As an exemplary dataset, countries with $N_{Country} > 500$ respondents were selected, then $N_{Country} = 500$ respondents from each of these countries were sampled. The resulting sample comprised $N_{Total} = 2,000$ respondents from Australia, Canada, Great Britain, and USA.

The model with a sum-to-zero constraint fills a gap between two existing modeling approaches. In contrast to the random threshold model it allows for dependencies between varying thresholds, but does not enforce a perfect negative correlation as is the case when ERS or MRS are specified a priori in a multidimensional PCM (see Figure 1). Therefore, the novel model is contrasted to these two approaches in the empirical analysis. Hence, the novel sum-to-zero model, a multidimensional PCM with symmetric thresholds shifts through a priori specified response style dimensions ERS and MRS, and a random threshold model with independent varying thresholds were fitted to the Big Five data. As a baseline, additionally, a PCM ignoring response styles was fitted.

As the data contained item responses from four countries, a multi-group setup was used. Item-specific threshold parameters $(-\beta_i + \tau_{ik})$ were set equal between countries which allows for the estimation and comparison of country-specific latent means and (co-)variances of Big Five and response style dimensions. Australia served as a reference country, and the means of the latent content trait and response

style dimensions were fixed to $\boldsymbol{\mu}_{Australia} = \mathbf{0}$; latent means were estimated for Canada, Great Britain, and USA. Variances and covariances for all four countries were estimated freely where applicable. They were fixed to 0 between varying thresholds and between varying thresholds and content traits for the random threshold model (Wang et al., 2006), and fixed to -1 between the outer and inner thresholds the multidimensional PCM to model ERS and MRS, respectively. The *R* package *TAM* (Kiefer et al., 2017; R Core Team, 2019) with Marginal Maximum Likelihood Estimation using a Quasi Monte-Carlo Integration procedure was used to estimate the models (see Appendix A).

Model Fit

Table 1 shows relative and absolute model fit indices for the multi-group analysis. Likelihood-Ratio tests compare the response style models to a PCM ignoring response styles. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two relative fit measures, where a smaller value indicates a better fit. The BIC includes a penalty for model complexity. The Standardized Generalized Dimensionality Discrepancy Measure (SGDDM) is a measure of absolute model fit in the metric of correlations, where a value close to 0 indicates the absence of local dependence and a value approaching 1 indicates the presence of local dependence.

Incorporating response styles into the IRT model increased model fit for all models compared to the PCM (all $p < .001$). The comparison of model fit between response style models is less evident: while the model with sum-to-zero constraint fits best in terms of AIC, the multidimensional PCM has the lowest BIC, and the

TABLE 1: Model Fit in the Multi-Group Analysis

	Response Style Dimensions	LR-Test	AIC	BIC	SGDDM
PCM	none	—	261,519	263,054	0.055
Sum-to-Zero	$K-1$ varying thresholds	$\chi^2(93) = 8,296$ $p < .001$	253,409	255,465	0.053
Multidimensional PCM	A priori specified ERS and MRS dimensions	$\chi^2(58) = 8,201$ $p < .001$	253,435	255,294	0.053
Random Threshold Model	K independent varying thresholds	$\chi^2(28) = 6,601$ $p < .001$	254,974	256,665	0.052

Note. PCM: Partial Credit Model, ERS: Extreme Response Style, BIC: Bayesian Information Criterion, SGDDM: Standardized Generalized Dimensionality Discrepancy Measure, LR-Test: Likelihood Ratio Test (comparison to the PCM).

random threshold model has the best absolute fit based on SGDDM. Overall, model fit improved with the inclusion of response styles, but there were no substantial empirical differences between response style models.

Hence, the appropriate model may be chosen by means of its informative value for the specific research scenario. For example, in case that one is certain that ERS and MRS are present in the data, a multidimensional PCM is a parsimonious model choice. In contrast, when one assumes that response tendencies may be unsystematic across respondents a random threshold model may rather be chosen. The novel model with sum-to-zero constraint is an appropriate choice when little is known about the type of response styles: it can account for ERS and MRS, but also for more uncommon response patterns in the data. I will elaborate on model parameters in the sum-to-zero model as well as their interpretation in the following.

Estimated Means, Variances, and Correlations

Figure 3 shows the estimated means and variances for each of the four countries in the varying threshold model using a sum-to-zero constraint. Mean differences between countries are negligible for Big Five Traits and varying thresholds (left panel). While country differences for variance estimates are negligible (right panel), there are substantial differences between variances of varying thresholds. Thresholds 1 and 4 have the largest variances, while thresholds 2 and 3 have

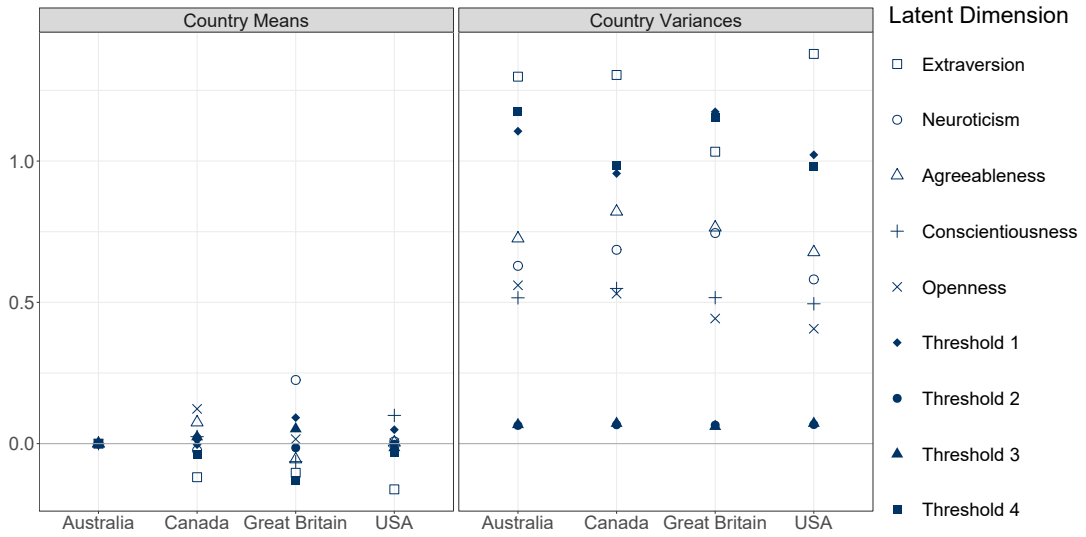

FIGURE 3: Estimated means (left) and variances (right) of the Big Five dimensions and varying thresholds (filled shapes) in the model with sum-to-zero constraint for each of the four countries (Australia served as a reference country).

variances close to zero. This result is in line with the good relative fit of the multidimensional PCM with ERS and MRS: a consistent finding is that the ERS trait has the largest variance among response style traits (see e.g., Plieninger & Heck, 2018; Wetzel, Böhnke, & Rose, 2016)[2].

The type of response styles in the data can be described by the estimated correlations between varying thresholds. As outlined above, the absolute height of correlations can only be interpreted with caution due to the ipsatization of varying thresholds (Clemans, 1966). However, we can interpret the relations between varying thresholds in terms of their rank order (Rost, 2004).

A strong negative correlation between the outer thresholds (1 and 4) would indicate ERS, a strong negative correlation between the inner thresholds (2 and 3) would indicate MRS (see Figure 1). Figure 4 shows the estimated correlations between varying thresholds in the model with sum-to-zero constraint, where the correlation between the outer and between the inner thresholds are displayed by filled shapes. We see a strong negative correlation between the outer thresholds, indicating ERS, but a weak correlation between the inner thresholds. The remaining threshold correlations are close to zero. Again, threshold correlations only differ marginally between countries.

Taken together, this pattern speaks in favor of a strong ERS tendency in the data. The presence of ERS would also explain the poor relative model fit of the random threshold model that cannot account for ERS as it restricts all varying threshold to be uncorrelated ($\mathbf{\Sigma} = Diag$). However, with the varying threshold

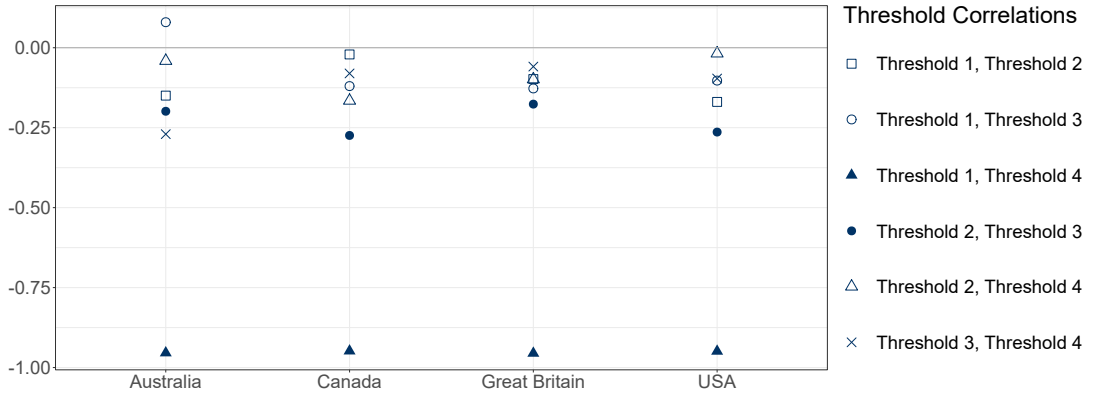

FIGURE 4: Estimated correlations between varying thresholds for the model using a sum-to-zero constraint. Filled shapes depict correlations between varying thresholds reflecting Extreme Response Style and Mid Response Style.

[2]Similarly, in the multidimensional PCM, variances of the thresholds were estimated in the same range between countries ($0.94 < Var^{ERS} < 1.24$; $0.07 < Var^{MRS} < 0.09$). Also in the random threshold model, the variance of the outer thresholds was larger than of the intermediate thresholds ($0.67 < Var^{T1;T4} < 0.94$; $0.05 < Var^{T2;T3} < 0.12$).

model using a sum-to-zero constraint, we can account for such negative correlations between thresholds and represent ERS in the psychometric model leading to an improved relative model fit.

Illustration of Threshold Shifts for Four Respondents

Figure 5 illustrate response patterns and category probability curves of four exemplary respondents under a sum-to-zero model. The leftmost column displays a respondent with threshold shifts close to expectation. The second column displays a respondent with negative ERS tendency: the respondent avoids the extreme categories and this is reflected by strong outward shifts of the outer thresholds. The third column displays a respondent with a preference for the first agreement category. This response pattern is reflected by outwards shifts of the threshold bounding this category. The rightmost column displays a respondent with unsystematic category preferences. He or she has a higher probability to respond in the middle category and prefers the highest agreement category over the moderate agreement category. The two exemplary respondents on the right-hand side of Figure 5 illustrate response patterns that cannot be captured in a multidimensional PCM with ERS and MRS as the combinations of threshold shifts

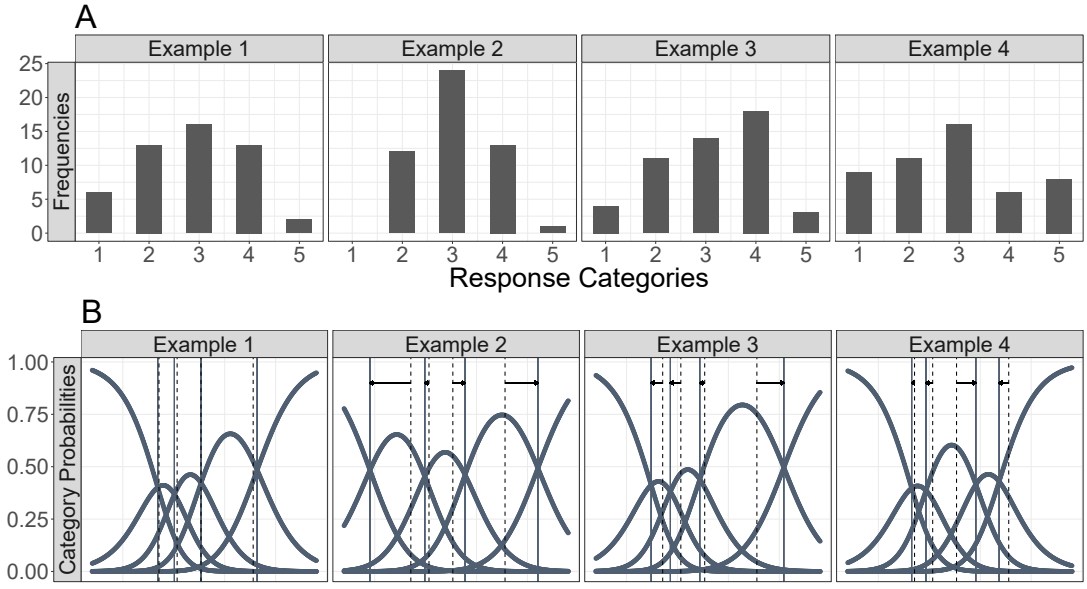

FIGURE 5: Frequency of category choices (A) and category probabilities (B) for four exemplary respondents; from left to right: respondent with small threshold shifts, respondent avoiding the extreme categories, respondent with a preference for the first agreement category, respondent who prefers the middle category, and the highest over the first agreement category.

differ from the specification of extreme or mid responding (see Figure 1), but can be accounted for in the varying threshold model with sum-to-zero constraint.

To conclude, it seems essential to account for response styles in the multi-country personality data, as model fit substantially increased when response styles were incorporated into the psychometric models. The empirical differences between the response style models remain inconclusive. ERS seems to be the dominant response style, but also further and less dominant response tendencies seem to be present in the data. In contrast to a multidimensional PCM that specifies response styles a priori, the varying threshold model with sum-to-zero constraint allows to accommodate different types of response styles, also initially unknown types. But it can also, in contrast to a random threshold model, account for consistently encountered response tendencies such as ERS or MRS. Most importantly, the novel model allows to test what kind of response tendencies are dominant in rating data, and whether there exist country-specific differences in response tendencies.

Discussion

The new model uses a sum-to-zero constraint on varying thresholds ($\sum_{k=1}^{K} \delta_{nk} = 0$) to separate content trait from response style effects. Through the inclusion of varying thresholds into the psychometric model, a large variety of response tendencies can be accounted for. This includes response styles such as ERS and MRS that imply symmetric threshold shifts around the item location, but also more individualized, unknown response tendencies. The sum-to-zero constraint allows to estimate the covariances between $K-1$ varying thresholds and between content trait and $K-1$ varying thresholds. The variance and covariances for the K^{th} threshold can be derived through a conversion of the estimated variances and covariances. A simulation study demonstrated that the model can validly estimate content trait, response style, and item-category parameters under various data conditions. Furthermore, a multi-country analysis using data of the Big Five personality factors showed that the model captures ERS as the dominant response style in the data, but also individual response tendencies that were unmodeled before.

The model with sum-to-zero constraint closes a gap between models that specified response styles a priori and imposed strong restrictions on varying thresholds (e.g., multidimensional PCMs, see Bolt & Johnson, 2009; Falk & Cai, 2016; Wetzel & Carstensen, 2017) on the one hand, and models that restricted varying thresholds to be uncorrelated (e.g., Wang et al., 2006) on the other hand. Furthermore,

restricting varying thresholds to sum to zero is a theoretically motivated constraint. As varying thresholds are centered within respondents, they reflect the dispersion of thresholds around the location of the respondent on the latent continuum, excluding a shift in location through response styles. Therefore, varying thresholds reflect respondents' perception of category width and do not increase or decrease item difficulty.

A remark must be made on the modeling of acquiescence. A prominent approach to account for ARS is the addition of an ARS parameter to the linear parameter combination for the agreement categories (Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Wetzel & Carstensen, 2017). For positive levels of ARS, the ARS trait parameter increases the probability for the agreement categories while decreasing the probabilities for the disagreement categories. But at the same time, it influences the location of the respondent on the latent continuum as agreement with the item becomes more probable for positive ARS trait levels (see Plieninger & Heck, 2018, for a discussion). When varying thresholds are constrained to sum up to zero—fixing the location of the respondent on the latent scale—such shift processes for ARS cannot be accounted for by the model. Accommodating ARS as an additional response style dimension in a varying threshold model may be an interesting topic for future research.

Other model extensions, which use generalized versions of IRT models with discrimination parameters for trait and response style dimensions, allow us to investigate the differential impact of latent dimensions on single items (e.g., Falk & Cai, 2016; Wang & Wu, 2011). The sum-to-zero constraint can also be extended to a generalized multidimensional PCM, where we have two sets of discrimination parameters: one for the content traits, and one for varying thresholds. In this case, item-specific discrimination parameters may be restricted to be equal for all varying threshold dimensions, and therewith indicate the impact of response tendencies on specific item responses. Future studies may assess the ability of such a model to capture differential influences of the latent content and response tendency dimensions in item responses.

The new varying threshold model may serve as a tool to further investigate processes underlying rating scale responses. As minimal a priori assumptions on response styles are used as constraints in the model, it is well suited to investigate response tendencies when little is known about how they might manifest themselves in rating data. Even though in the multi-country comparison presented here, country differences were negligible, this may not be the case for other country samples (see e.g., Bolt et al., 2014; G. W. Cheung & Rensvold, 2000). In particular

when response style types are unknown, a flexible model capturing the individual type of response tendencies for each country is a valuable tool for such cross-country comparisons.

To conclude, the approach proposed here extends the literature of IRT response style models. By using the sum-to-zero constraint on threshold variations, it accounts for individual differences in response tendencies without imposing strict assumptions on the type of response style that is modeled. Hence, content trait and the perception of the rating scale of the respondent that are usually entangled in the rating responses can be separated in a psychologically meaningful way.

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*, 1–20. doi:10.1111/bmsp.12169

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics. *R package version 2.3*. Retrieved from https://cran.r-project.org/package=gridExtra

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

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

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

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

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

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

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

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

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

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

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika*, *30*, 99–121. doi:10.2333/bhmk.30.99

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

Cheung, M. W.-L. (2004). A direct estimation method on analyzing ipsative data with Chan and Bentler's (1993) method. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*, 217–243. doi:10.1207/s15328007sem1102_5

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. Richmond, VA: Psychometric Society. Retrieved from https://www.psychometricsociety.org/sites/default/files/pdf/MN14.pdf

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational and Organizational Psychology*, *67*, 89–100. doi:10.1111/j.2044-8325.1994.tb00553.x

Costa, P. t., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Metthews, & D. H. Saklofske (Eds.), *The SAGE Handbook of Personality Theory and Assessment: Personality Measurement and Testing* (2nd). London: SAGE Publications Ltd.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*, *29*, 115–126. doi:10.1207/s15327906mbr2901_4

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

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

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

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management*, *6*, 243–266. doi:10.1177/1470595806066332

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

Kelley, K. (2018). MBESS: The MBESS R Package (R package version 4.4.3). Retrieved from https://cran.r-project.org/package=MBESS

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

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

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

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

Mersmann, O., Trautmann, H., Steuer, D., & Bornkamp, B. (2018). truncnorm: Truncated Normal Distribution (R package version 1.0-8). Retrieved from https://cran.r-project.org/package=truncnorm

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

Müller, K. (2017). here: A simpler way to find your files (R package version 0.1). Retrieved from https://cran.r-project.org/package=here

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

Open Source Psychometrics Project. (2019). Open psychology data: Raw data from online personality tests. Retrieved from https://openpsychometrics.org/_rawdata/

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

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

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

Radcliffe, J. A. (1963). Some properties of ipsative score matrices and their relevance for some current interest tests. *Australian Journal of Psychology*, *15*, 1–11. doi:10.1080/00049536308255468

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

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion.* (2nd Edition). Bern, Göttingen, Toronto, Seattle.

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

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

Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

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., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, *27*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

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

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (Chap. Res, pp. 349–363). doi:10.1093/med:psych/9780199356942.003.0024

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

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

Wetzel, E., 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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A grammar of data manipulation. *R package version 0.7.6*. Retrieved from https://cran.r-project.org/package=dplyr

Appendix A

Implementation in Standard Software

The varying threshold model with sum-to-zero constraint can be formulated as a multidimensional Partial Credit Model (PCM) model for polytomous responses. Herein, each varying threshold can be formalized as an additional person dimension with fixed scoring weights that reflect threshold shifts that sum to zero within respondents (see Thissen & Steinberg, 1986).

The scoring weights for varying thresholds can be derived through cumulating influence of varying thresholds across categories (see Equation 4). Table A1 shows the threshold and category probability functions of a model with varying thresholds with $K = 4$. In order to implement the sum-to-zero constraint $\sum_{k=1}^{K} \delta_{nk}$, a restriction on δ_{n4} is imposed so that $\delta_{n4} = -\delta_{n1} - \delta_{n2} - \delta_{n3}$.

TABLE A1: Threshold and Category Probability in a Varying Threshold Model Using a Sum-to-Zero Constraint on δ_{nk} for an Item With $K + 1$ Response Categories with $k \in \{0, ..., 4\}$ and $\delta_{n4} = -\delta_{n1} - \delta_{n2} - \delta_{n3}$

k	Threshold Probability	Category Probability
0		$\frac{\exp(0)}{C}$
1	$\frac{\exp(\theta_n-\beta_i-\tau_{i1}+\delta_{n1})}{1+\exp(\theta_n-\beta_i-\tau_{i1}+\delta_{n1})}$	$\frac{\exp(1\cdot\theta_{n1}-\beta_i-\tau_{i1}+\delta_{n1})}{C}$
2	$\frac{\exp(\theta_n-\beta_i-\tau_{i2}+\delta_{n2})}{1+\exp(\theta_n-\beta_i-\tau_{i2}+\delta_{n2})}$	$\frac{\exp\left(2\cdot\theta_{n1}-\sum_{k'=0}^{2}(\beta_i+\tau_{ik'})+\delta_{n1}+\delta_{n2}\right)}{C}$
3	$\frac{\exp(\theta_n-\beta_i-\tau_{i3}+\delta_{n3})}{1+\exp(\theta_n-\beta_i-\tau_{i3}+\delta_{n3})}$	$\frac{\exp\left(3\cdot\theta_{n1}-\sum_{k'=0}^{3}(\beta_i+\tau_{ik'})+\delta_{n1}+\delta_{n2}+\delta_{n3}\right)}{C}$
4	$\frac{\exp(\theta_n-\beta_i-\tau_{i4}-\delta_{n1}-\delta_{n2}-\delta_{n3})}{1+\exp(\theta_n-\beta_i-\tau_{i4}-\delta_{n1}-\delta_{n2}-\delta_{n3})}$	$\frac{\exp\left(3\cdot\theta_{n1}-\sum_{k'=0}^{4}(\beta_i+\tau_{ik'})+\delta_{n1}+\delta_{n2}+\delta_{n3}-\delta_{n1}-\delta_{n2}-\delta_{n3}\right)}{C}$

Note. C is a normalizing constant so that the category probabilities sum up to 1: $\sum_{k'=0}^{K} \exp(s_{k'}\theta_n + \sum_{k^*=0}^{k'}(-\beta_i - \tau_{ik^*} + \delta_{nk^*})$ and $\theta_n - \beta_i - \tau_{i0} + \delta_{n0} \equiv 0$ for identification; n for persons, i for items, k for thresholds, θ_n for person parameters, β_i for item parameter, τ_{ik} for threshold parameters, and δ_{nk} for thresholds varying between respondents

From Table A1, we can see that δ_{n1} impacts all subsequent Categories 1 to 3, while δ_{n2} impacts all subsequent Categories 2 to 3, etcetera. The total impact of δ_{n1}, δ_{n2}, and δ_{n3} on Category 4 is zero, as varying thresholds are first added through the cumulation across categories, but then subtracted through the sum-to-zero restriction ($\delta_{n4} = -\delta_{n1} - \delta_{n2} - \delta_{n3}$).

We can express δ_{nk} as additional trait parameters θ_n^{δ} with their influence on single categories being described by scoring weights $\boldsymbol{s}_k^{\delta}$. Thus, we can reparameterize the model from Equation 4 into a multidimensional PCM with $K-1$ additional

trait dimensions for varying thresholds (here 3 additional dimensions):

$$p(X_{ni} = k) = \frac{\exp\left(s_k\theta_n - \sum_{k'=0}^{k}(\beta_i + \tau_{ik'}) + s_k^{\delta_1}\theta_n^{\delta_1} + s_k^{\delta_2}\theta_n^{\delta_2} + s_k^{\delta_3}\theta_n^{\delta_3}\right)}{\sum_{j=0}^{K}\exp\left(s_j\theta_n - \sum_{k'=0}^{j}(\beta_i + \tau_{ik'}) + s_j^{\delta_1}\theta_n^{\delta_1} + s_j^{\delta_2}\theta_n^{\delta_2} + s_j^{\delta_3}\theta_n^{\delta_3}\right)}. \quad (8)$$

The information from Table A1 can then be used to derive scoring weights s_k^{δ} for the $K-1$ varying threshold dimensions θ_n^{δ} (Table A2). As δ_{n4} is a function of δ_{n1}, δ_{n2}, and δ_{n3} only three latent traits are estimated and through the sum-to-zero constraint, the effect of each trait on the last category is 0 (see Table A1).

TABLE A2: Scoring Weights for Content Trait and Varying Thresholds δ_{nk} for 5 Response Categories With $k \in \{0, ..., 4\}$ Using a Sum-to-Zero Constraint

	Cat. 0	Cat. 1	Cat. 2	Cat. 3	Cat. 4
$\boldsymbol{s}$	0	1	2	3	4
$\boldsymbol{s}_{reversed-coded}$	4	3	2	1	0
$\boldsymbol{s}^{\delta_1}$	0	1	1	1	0
$\boldsymbol{s}^{\delta_2}$	0	0	1	1	0
$\boldsymbol{s}^{\delta_3}$	0	0	0	1	0

The formulation of the varying threshold as a multidimensional PCM allows us to estimate the model in standard software for multidimensional IRT models such as Mplus (Muthén & Muthén, 2012) or the *R* programing environment (R Core Team, 2019) with the package *TAM* (Kiefer et al., 2017) or *mirt* (Chalmers, 2012) using Marginal Maximum Likelihood estimation. Exemplary code to fit a model on rating data with ten items (of which 5 are reversed-coded), five response categories ($k \in \{0, ..., K\}$), and three varying thresholds in *R* with the package *TAM* and a Quasi Monte-Carlo Integration procedure may be:

```
# An Overview of the Data
> head(dat)
  item1 item2 item3 item4 item5 item6 item7 item8 item9 item10
1     2     1     1     1     1     1     2     3     2      3
2     1     1     2     1     1     0     1     1     1      3
3     1     1     1     1     1     2     3     3     3      3
4     2     2     2     4     2     0     1     0     0      2
5     3     3     3     3     3     1     1     0     2      1
6     2     0     2     0     2     1     3     4     3      3

# Define Fixed Numbers
nItems <- 10 # Number of Items
whichRev <- 6:10 # which Items are Reversed-Coded
nCat <- 5 # Number of Categories
nThres <- nCat - 1 # Number of Thresholds
nDimContent <- 1 # Number of Content Dimensions
nDimTotal <- 1 + nThres - 1 # Total Number of Dimensions

# Create Design Matrix B for Multidimensional PCM
B <- array(0, dim = c(nItems,nCat,nDimTotal)) # I x (K+1) x D

# Fill in Weights for Content Trait
B[,,1] <- rep(0:4, each = nItems)
B[whichRev,,1] <- rep(4:0, each = length(whichRev)) # Reversed-Coded

# Fill in Weights for Varying Thresholds
B[,,2] <- rep(c(0,1,1,1,0), each = nItems) # Threshold 1
B[,,3] <- rep(c(0,0,1,1,0), each = nItems) # Threshold 2
B[,,4] <- rep(c(0,0,0,1,0), each = nItems) # Threshold 3

# Model Fit
fit_SumToZero <- TAM::tam.mml(resp=dat,
                              B=B,
                              control = list(snodes=5000,
                                             maxiter = 5000))
```

Different Styles, Different Times: How Response Times can Inform our Knowledge About the Response Process in Rating Scale Measurement

Mirka Henninger and Hansjörg Plieninger

University of Mannheim

Abstract

When respondents use different ways to answer rating scale items, they employ so-called response styles that bias inferences drawn from measurement. To describe the influence of such response styles on the response process, we investigated relations between extreme, acquiescent and mid response style and response times in three datasets using multilevel modeling. On the response level, agreement and midpoint, but not extreme responses were slower. On the person level, response times increased for extreme, but not for acquiescence or mid response style traits. For all three response styles, we found negative cross-level interaction effects, indicating that a response matching the response style is faster. The results demonstrate that response styles facilitate the choice of specific category combinations in terms of response speed across a wide range of response style trait levels.

Rating scales are often used to measure latent variables such as beliefs, attitudes or personality traits as they are convenient to apply and evaluate. However, the response to a rating scale item does not only reflect the trait to be measured, but also the way a respondent perceives and uses the rating scale. The so-called response styles (Paulhus, 1991) can be regarded as latent traits that describe the respondents' tendencies to prefer certain types of categories over others irrespective of item content. For example, a bias towards choosing the highest and lowest categories is called *extreme response style* (ERS), a tendency to generally agree with the item is called *acquiescence response style* (ARS), and a preference towards the middle category is called *mid response style* (MRS; see Van Vaerenbergh & Thomas, 2013, for a review and definitions of additional response styles).

Response styles seem to be ubiquitous in rating data (e.g., Böckenholt & Meiser, 2017; Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel, Carstensen, & Böhnke, 2013). Moreover, response styles have been shown to be consistent across different content traits (Weijters, Geuens, & Schillewaert, 2010a; Wetzel et al., 2013), and to be stable personality characteristics that persist over time (Weijters, Geuens, & Schillewaert, 2010b; Wetzel, Lüdtke, Zettler, & Böhnke, 2016). Thus, rating scales do not only capture information on the latent content trait, but also on response styles. Such response styles distort trait measurement precision (Bolt, Lu, & Kim, 2014; Wetzel & Carstensen, 2017), inflate relations between variables (Abad, Sorrel, Garcia, & Aluja, 2018; Böckenholt & Meiser, 2017), or bias cross-group comparisons, for example in cross-cultural research (Bolt et al., 2014; Rollock & Lui, 2016).

Attempts to explain response styles through demographic, personality, and situational variables yielded mixed results. The effects of gender and age on ERS are inconsistent across studies (e.g., Hamilton, 1968; Moors, 2008; Van Vaerenbergh & Thomas, 2013; Weijters, Geuens, & Schillewaert, 2010b), but intelligence, occupational status and education seem to reduce ERS (e.g., Bolt & Johnson, 2009; Meisenberg & Williams, 2008). On the one hand, ERS increases with certain personality traits, such as intolerance of ambiguity, simplistic thinking, and decisiveness (Naemi, Beal, & Payne, 2009), on the other hand, the relation of response styles and the Big Five have been found to be positive, negative, or non-existent (e.g., Austin, Deary, & Egan, 2006; Couch & Keniston, 1960; Grimm & Church, 1999; He & Van De Vijver, 2013; Hibbing, Cawvey, Deol, Bloeser, & Mondak, 2017; van Dijk, Datema, Piggen, Welten, & van de Vijver, 2009; Wetzel & Carstensen, 2017). Situational variables, such as reducing the number of response categories and inducing cognitive load increases the magnitude of ERS and ARS,

respectively (Cabooter, 2010; Knowles & Condon, 1999; Weijters, Cabooter, & Schillewaert, 2010), while at the same time alternative response formats have been shown to reduce, but also increase response styles (Böckenholt, 2017; Plieninger, Henninger, & Meiser, 2019).

The inconsistent results with respect to personality and situational covariates demonstrate how little is still known about response styles as a psychological phenomenon in the nomological net. Hence, response styles need to be investigated and response times may be a means to this end. Fekken and Holden (1994) argued that the time respondents take to provide a self-report response is a behavioral representation of the underlying cognitive process. They showed that response times are meaningful indicators for the trait to be measured on a personality test. Since responses are not only indicators of the trait to be measured but also of response styles, the time accompanying the responses should also be an indicator of processes related to content as well as response styles. Knowledge about the cognitive processes that influence response category selection through response styles will help us to evaluate the often made claim that response styles are a result of reduced cognitive effort (e.g., Aichholzer, 2013; Krosnick, 1999), and to evaluate the magnitude of impact that response styles have on data quality.

Response Times in Rating Scale Measures

Response times have been used to assess cognitive processes in experimental psychology (e.g., Heck & Erdfelder, 2016) and served as collateral information in IRT models for ability testing (e.g., van der Linden, Klein Entink, & Fox, 2010). However, there is little research investigating response times in personality measurement and even fewer assessing the relationship of response times and response styles.

Response Times in Personality Measurement

Response times have served as an indicator of respondents' motivation and deliberation in surveys. Fast responses have been associated with low motivation of the respondent (Callegaro, Yang, Bhola, Dillman, & Chin, 2009), lower validity (Neubauer & Malle, 1997) and poor data quality (Zhang & Conrad, 2013). Furthermore, items that appear later in the survey are responded faster than earlier items and with a lower variability in the responses, which might be an indicator of decreasing motivation of respondents towards the end of the survey (Callegaro et al., 2009; Galesic & Bosnjak, 2009; Wise & DeMars, 2005, 2006;

Yan & Tourangeau, 2008). Similarly, while shorter response times are associated with reports of desirable attitudes and behavior, longer response times have been linked to responses that are given more carefully, such as faked responses or the reporting of undesirable attitudes (Andersen & Mayerl, 2017; Dunn, Lushene, & O'Neil, 1972; McIntyre, 2011; Neubauer & Malle, 1997; van Hooft & Born, 2012).

Another view on response times links fast responses to high confidence in the rating. Fast responses have been associated with the accessibility of the trait being measured, as respondents whose attitudes were important to them responded faster (Tourangeau, Rasinski, & D'Andrade, 1991). Similarly, fast response times are associated with a high consistency in item responses since respondents take less time to decide for a response option when they are certain about it (McIntyre, 2011). In line with that, slow responses are considered to indicate cognitive effort in the response process. When respondents try to find the best answer to the item, response times increase, especially for complicated or ambiguous questions (Bassili & Scott, 1996; Dunn et al., 1972; Hanley, 1965; Rogers, 1973). Similarly, item complexity such as the number of clauses, characters, or cognitive operations required for a response increases response times (Kulas & Stachowski, 2009; Lenzner, Kaczmirek, & Lenzner, 2010; Sauer, Auspurg, Hinz, & Liebig, 2011; Yan & Tourangeau, 2008).

In sum, fast responses can have two interpretations: they may indicate a spontaneous response mode, in which respondents demonstrate low motivation and deliberation, but may also indicate confidence in the rating as the optimal response is highly accessible. Slower responses are the result of a careful, effortful or deliberate cognitive process, either due to thought-out decisions or item complexity.

Response Times in Response Style Research

In this research project, we examine the relation between extreme, acquiescent, and mid responding and response times to describe cognitive processes in rating scale usage. Herein, we differentiate between specific *responses* (e.g., extreme, agree, or mid responses) that are given faster or slower than other responses, and *respondents* (with different ERS, ARS, or MRS levels) that may respond faster or slower than other respondents across items.

Effects of Current Responses on Response Times at the Response Level

In terms of extreme responses, the results of Casey and Tryon (2001) showed that a majority of participants gave faster responses in the extreme categories than in the neighboring non-extreme category. This result may suggest a negative main effect of extreme responses on response times, but the stability and magnitude of the effect remains unclear.

> Hypothesis 1a) Extreme responses may result in shorter response times (although evidence for this effect is based on only one investigation by Casey & Tryon, 2001).

Agree responses might be related to task complexity and a result of cognitive burden when items are hard to interpret. Agreement to both reversed and non-reversed items occurs with complex rather than easy items and results in higher cognitive demand and longer response times (Hanley, 1965; Rogers, 1973; Swain, Weathers, & Niedrich, 2008). In addition, Knowles and Condon (1999) showed in an experimental investigation that under high cognitive load, respondents tended to agree with the items more often. As cognitive load has been associated with longer response times, this effect further supports the hypothesis that agree responses lead to longer response latencies.

> Hypothesis 1b) We expect that agree responses are given slower than non-agree responses since agree responses have been shown to result in longer response times (Hanley, 1965; Knowles & Condon, 1999; Rogers, 1973; Swain et al., 2008). Slower agree responses may indicate task complexity and increased cognitive demand.

Kulas and Stachowski (2009) found that respondents took longest to give a response in the middle category. The authors argued that it is cognitively less demanding to agree or disagree than to choose the midpoint. Especially when respondents cannot decide for a directed response, the choice of the undecisive midpoint may indicate a well evaluated, and therefore cognitively demanding judgment process that becomes visible through response times.

> Hypothesis 1c) Mid responses may take longer than directed responses based on the evidence and considerations presented by Kulas and Stachowski (2009). Similar to the process underlying ARS, slower responses may indicate cognitive burden in evaluating the item, leading to a thought-out item response.

Effects of Response Style Traits on Response Times at the Respondent Level

As there is little evidence pointing towards directed effects for ERS, ARS, and MRS on response times, hypotheses on the respondent level are exploratory. First, evidence for ERS is mixed. On the one hand, fast respondents showed higher variability in their responses than slow respondents (Neubauer & Malle, 1997). High variability in the responses is associated with high ERS levels, as the variance in the responses increases when extreme categories are chosen more often which may be indirect evidence that high ERS trait levels reduce response times. On the other hand, Naemi et al. (2009) found no main effect of ERS levels on response times.

> Exploratory Analysis 2a) We will explore the effects of ERS trait levels on response times. As the effects reported by Neubauer and Malle (1997) are indirect, and no effects were found by Naemi et al. (2009), no prediction can be made on whether high ERS levels should lead to shorter, faster, or unchanged response times.

For ARS, Mayerl (2013) argued that measured attitudes are stronger influenced by acquiescence when respondents answered in a fast, automatic-spontaneous response mode. In line with that, the descriptive response times by Knowles and Condon (1999) indicate lower response times for respondents with high ARS levels than for respondents with low ARS levels.

> Exploratory Analysis 2b) We will explore the effects of ARS trait levels on response times. First results (Knowles & Condon, 1999; Mayerl, 2013) point towards a decrease in response times for higher ARS trait levels, but overall evidence is sparse.

To our knowledge, there is no literature to build on in order to predict effects of MRS levels on response times.

> Exploratory Analysis 2c) We will explore the effects of response style trait levels for MRS on response times.

Interaction Effects Between the Current Response and Response Style Traits on Response Times

Besides main effects of item responses and respondents' response style traits, interaction effects may occur such that respondents with higher response style

traits are faster when they give responses matching their response style trait. For example, a respondent with high ERS trait levels may be faster when giving an extreme response, and slower when giving a non-extreme response.

In terms of ERS, Naemi et al. (2009) showed that the combination of ERS and specific personality traits jointly decrease response times. This pattern, speaks in favor of a more complex relation between ERS and response times.

In terms of ARS, Knowles and Condon (1999) found an interaction effect in such a way that respondents with high levels of ARS were faster when they agreed than when they disagreed with an item, and faster when they agreed than non-ARS respondents.

For MRS, there is no literature directly pointing towards an interaction effect for MRS and response times. However, response time for choices of the mid response option may be longer for respondents that have weighed the pros and cons of either side of the item, but that do not have a general tendency to prefer the middle category over the other response options. In contrast, respondents with a high MRS trait, using the mid response option abundantly may have faster response times when giving a mid response than respondents with low MRS trait levels (see Kulas & Stachowski, 2009).

Speed-Distance Hypothesis

An important theory that further supports the idea of an interaction effect between response style traits and item responses on response times is the speed-distance hypothesis. It predicts that that response times decrease with increasing distance between the trait level of the respondent and item difficulty (Akrami, Hedlund, & Ekehammar, 2007; McIntyre, 2011). Larger distances result in a higher confidence to give a clear-cut response, while smaller distances imply high uncertainty about the item response (see also Ferrando & Lorenzo-Seva, 2007; Ranger & Ortner, 2011, for two IRT models based on the speed-distance relationship).

Evidence for the speed-distance hypothesis is abundant. For example, Fekken and Holden (1992) showed that response times for respondents with high trait levels that agree with the item respond fast, while respondents with high trait levels that disagree with the item respond slowly. Similarly, Casey and Tryon (2001) and McIntyre (2011) argued that pronounced self-schemata guide responses and decrease response times. In contrast, respondents with low trait knowledge or respondents that answer contrary to their self-schemata give slow responses (see also Dunn et al., 1972; Kuiper, 1981). The complex relationship between the trait level, the given response and response times even holds for peer ratings. Fuhrman

and Funder (1995) found that high self-ratings were predictive of higher as well as quicker peer ratings; peer ratings were slower when the trait was rated high, but the current item was disagreed with. In short, the speed-distance hypothesis assumes that the more likely a response, the faster it will be given.

Based on the speed-distance hypothesis, we predict that the closer the observed response matches the response style trait, the faster the response will be. For example, a person with high ERS levels will take little time to give an extreme response. In contrast, when deviating from his or her ERS trait by giving a non-extreme response, the respondent will take more time. This reasoning is also in line with the evidence that, under high confidence, responses are given faster (McIntyre, 2011; Tourangeau et al., 1991) while responses involving high cognitive effort are given slower (e.g., Kulas & Stachowski, 2009; Lenzner et al., 2010; Sauer et al., 2011; Yan & Tourangeau, 2008).

> Hypotheses 3a-c) We predict for ERS, ARS, and MRS that responses that are in line with the response style traits will be given faster, whereas responses that are opposite to the response style trait will be accompanied by longer response times.

Method

Collecting Response Time Data

In collaboration with three research groups (Fladerer & Misterek, 2018; Pfister, 2018; Plieninger et al., 2019), we recorded response times for each response in three studies. The first study was conducted in collaboration with Pfister (2018) who initially investigated the relation between implicit personality measures and response styles in rating scale items with five categories. The second study originated from a collaboration with Plieninger et al. (2019), wherein the authors compared different response formats using six response categories; we collected response times in the Likert condition. The third study consists of responses to 5- and 7-point rating scales on Leadership and Team Collaboration and was conducted in collaboration with Fladerer and Misterek (2018).

When planning the three studies, we aimed at validly measuring response styles by using heterogeneous items without a common trait (see De Beuckelaer, Weijters, & Rutten, 2010; Greenleaf, 1992), while at the same time making the study conditions as close to real measurement situations as possible. Thus, we designed three studies accordingly: Study 1 focused on measurement of response

styles, therefore only heterogeneous items (i.e. items without a common trait) were selected from various scales (see De Beuckelaer et al., 2010; Greenleaf, 1992). Study 2 served as an intermediate step employing heterogeneous items as well as two content trait scales. Study 3 used items of five different, validated scales from organizational psychology to ensure that the results obtained from Study 1 and Study 2 can be generalized to applied measurement situations. Table 1 provides an overview of sample size, number of items, number of response categories, employed scales, and number of items per scale in each study[1]

TABLE 1: Overview of the Data Used for Analyses

	N	I	K	Scales
Study 1	161	39	5	*Heterogeneous* (no common trait; 39 items)
Study 2	154	54	6	Honesty-Humility (10 items) Personal Need for Structure (12 items) *Heterogeneous* (no common trait; 32 items)
Study 3	786	45	5	Identity Leadership Inventory (14 items)
			5	Social Identification (6 items)
			7	Perceived Organizational Support (8 items)
			7	Collective Self-Esteem (7 items)
			7	Resilience CD-RISC 10-item form (10 items)

Note. N: number of participants after exclusions; I: number of items, K: number of response categories

All three studies were conducted online. We used Javascript to track the response as well as the time in milliseconds associated with each mouse click. The response times for a given item was then operationalized as the time difference between the current and the preceding mouse click. For future research or applications, we made the Javascript code to collect response times available on OSF.

Data Preprocessing

Since data were collected online in the three studies, careful attention was paid to retain only valid data. The first two studies contained several validity checks to ensure data quality (e.g., items wherein participants could indicate that, for

[1]Honesty-Humility scale (Lee & Ashton, 2006); Personal Need for Structure scale (Machunsky & Meiser, 2006); Identity Leadership Inventory (Steffens et al., 2014); Social Identification scale (Mael & Asiforth, 1992); Perceived Organizational Support scale (Eisenberger, Huntington, Hutchison, & Sowa, 1986); Collective Self-Esteem scale (Riggs, Warka, Babasa, Betancourt, & S., 1994); Resilience CD-RISC scale (Sarubin et al., 2015). Heterogeneous item sets are provided on OSF: https://osf.io/gqb4y

example, they have been distracted during the study and one Bogus item, see also Meade & Craig, 2012). Based on the validity checks, we excluded 26 and 44 participants in Study 1 and 2, respectively. In the third study, respondents who answered to at least 30 out of 45 items were included in the analyses.

When participants were directed back to the preceding page because they omitted one or more items, we decided to exclude responses to the initially omitted items from analyses since they may be imprecise with respect to response times when respondents have to reorientate themselves on the survey page (see also Höhne & Schlosser, 2018)[2]. As we collected response times for each mouse click, we evaluated whether respondents answered to items more than once. Across all items and respondents, 9% of responses were changed in Study 1, 8% in Study 2, and 6% in Study 3. Assuming that a spontaneous response was the best indicator of the underlying response process, we used the initial response to an item in cases where participants later modified their response.

Response Times

Based on a Box-Cox transformation test, we log transformed response times to obtain a normal distribution. Assuming that very slow and very fast responses may not be the result of a valid response process, we excluded responses that deviated $+/-2$ SD from the individual respondent's mean response time (a common approach in response time analyses, see Bassili & Fletcher, 1991; Mayerl & Urban, 2008; Mulligan, Grant, Mockabee, & Monson, 2003). Through this procedure 302 out of $6,268$ responses (4.8%) in Study 1, 358 out of $8,279$ responses (4.3%) in Study 2 and $1,739$ out of $34,854$ responses (5.0%) in Study 3 were excluded which led to an approximately normal distribution of log response times on the sample level. Table 2 shows the descriptive sample statistics of log response times in the three datasets.

Table 2: Descriptive Sample Statistics of Log Response Times in the Three Datasets

	Min.	1st Quart.	Median	Mean	3rd Quart.	Max.
Study 1	-1.02	1.25	1.55	1.58	1.89	4.48
Study 2	-0.16	1.45	1.78	1.81	2.14	4.21
Study 3	-1.33	1.32	1.72	1.75	2.15	5.60

[2] 13 participants were redirected to a previous page in Study 1, 28 participants were redirected in Study 2. A majority of participants initially omitted one or two items on the previous page, two participants omitted all items of the Honesty-Humility scale. In Study 3, no participants were redirected to previous pages.

Response Style Indicators

We recoded item responses to obtain dichotomous response style indicators (see De Beuckelaer et al., 2010; Greenleaf, 1992; Wetzel & Carstensen, 2017). For extreme responses, a response was coded 1 if it was in either one of the two extreme categories and 0 otherwise. For agreement responses, responses in the agreement categories (i.e. categories above the mid point) were coded 1 and 0 otherwise. A response was coded a midpoint response with value 1, if the midpoint was chosen and 0 otherwise; midpoint responses were not defined in case of a scale with an even number of categories. Table 3 gives an overview of the scoring rules for scales with different numbers of response categories.

TABLE 3: Recoding of Item Responses Into Dichotomous Response Style Indicators for Different Number of Response Categories

		Initial Response						
Number of Categories	Response Type	-3	-2	-1	0	1	2	3
5	$X_{in}^{Extreme}$	-	1	0	0	0	1	-
	X_{in}^{Agree}	-	0	0	0	1	1	-
	X_{in}^{Mid}	-	0	0	1	0	0	-
6	$X_{in}^{Extreme}$	1	0	0	-	0	0	1
	X_{in}^{Agree}	0	0	0	-	1	1	1
	X_{in}^{Mid}	-	-	-	-	-	-	-
7	$X_{in}^{Extreme}$	1	0	0	0	0	0	1
	X_{in}^{Agree}	0	0	0	0	1	1	1
	X_{in}^{Mid}	0	0	0	1	0	0	0

Note. X_{in}: response to item i by person n for extreme, agree, and mid responses; no midpoint response was modeled for scales with an even number of response categories.

Multilevel Modeling Approach

We used a multilevel modeling approach to predict individual log response times based on responses of respondent n to item i using item responses (Level 1), respondents' response styles (Level 2) and their cross-level interaction as predictor variables.

On Level 1 (item response level), we used three dichotomous variables ($X_{in}^{Extreme}$, X_{in}^{Agree}, X_{in}^{Mid}) that indicated whether a given response was an extreme, agreement or midpoint response, respectively (see Table 3). Hence, Level 1 variables described whether the current item response was indicative of a specific response style. In addition, we entered effect-coded item fixed effects $\sum_{i=2}^{I} \beta_i X_i^{item}$ using X_1^{item} as a

reference to control for differences in response times due to item features, such as item length or complexity. Thus the Level 1 model equation is given by:

$$\begin{aligned}\text{log Response Times}_{in} = &\sum_{i=2}^{I} \beta_i X_i^{item} + \\ &\beta_{0n} + \beta_{1n} X_{in}^{Extreme} + \beta_{2n} X_{in}^{Agree} + \beta_{3n} X_{in}^{Mid} + \\ &e_{in}\end{aligned}$$

Level 2 (respondent level) variables were trait scores of response styles ERS, ARS, and MRS ($\theta_n^{ERS}, \theta_n^{ARS}, \theta_n^{MRS}$) for each respondent. The trait scores reflected interindividual differences in response styles. Rather than using, for example, manifest sum scores which may lead to biased estimates (see Lüdtke et al., 2008), we used a latent aggregation procedure. It takes sampling error into account when Level 1 variables ($X_{in}^{Extreme}, X_{in}^{Agree}, X_{in}^{Mid}$) are combined to form Level 2 variables ($\theta_n^{ERS}, \theta_n^{ARS}, \theta_n^{MRS}$). Therewith, we account for unreliability in Level 2 predictors and can correct for biases in between-group regression coefficients (see also Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Lüdtke et al., 2008; Marsh et al., 2009).

On Level 2, we specified a random intercept for respondents to account for differences in response times between respondents. We defined the parameters of the random intercept as a function of the latent response style traits θ_n^{ERS}, θ_n^{ARS}, and θ_n^{MRS}. The model equation for the intercept parameters is given by:

$$\beta_{0n} = \gamma_{00} + \gamma_{01}\theta_n^{ERS} + \gamma_{02}\theta_n^{ARS} + \gamma_{03}\theta_n^{MRS} + u_{0n}$$

Besides, we defined varying slope parameters $\beta_{1n}, \beta_{2n}, \beta_{3n}$ for each response type ($X_{in}^{Extreme}$, X_{in}^{Agree}, X_{in}^{Mid}) and defined them as a function of the respective latent response style trait ($\theta_n^{ERS}, \theta_n^{ARS}, \theta_n^{MRS}$), to study the effects of response styles on response times through cross-level interactions. The model equation for the slope parameters is given by:

$$\begin{aligned}\beta_{1n} &= \gamma_{10} + \gamma_{11}\theta_n^{ERS} + u_{1n} \\ \beta_{2n} &= \gamma_{20} + \gamma_{21}\theta_n^{ARS} + u_{2n} \\ \beta_{3n} &= \gamma_{30} + \gamma_{31}\theta_n^{MRS} + u_{3n}\end{aligned}$$

The resulting joint model equation is thus given by:

$$\begin{aligned}\text{log Response Times}_{in} = & \sum_{i=2}^{I} \beta_i X_i^{item} + \\ & \gamma_{00} + \gamma_{01}\theta_n^{ERS} + \gamma_{02}\theta_n^{ARS} + \gamma_{03}\theta_n^{MRS} + \\ & \gamma_{10}X_{in}^{Extreme} + \gamma_{11}\theta_n^{ERS}X_{in}^{Extreme} + \\ & \gamma_{20}X_{in}^{Agree} + \gamma_{21}\theta_n^{ARS}X_{in}^{Agree} + \\ & \gamma_{30}X_{in}^{Mid} + \gamma_{31}\theta_n^{MRS}X_{in}^{Mid} + \\ & u_{0n} + u_{1n}X_{in}^{Extreme} + u_{2n}X_{in}^{Agree} + u_{3n}X_{in}^{Mid} + e_{in}\end{aligned}$$

In summary, the model captures differences in response times due to simple interindividual differences (via β_{0n}) and difference due to item characteristics (via β_i). Thus, further effects can be interpreted as the deviation of the respondent's response time to an average item from his or her average response time. Main effects on Level 1 ($\gamma_{10}, \gamma_{20}, \gamma_{30}$) indicate whether specific responses (e.g., $X_{in}^{Extreme}$) take longer, main effects on Level 2 ($\gamma_{01}, \gamma_{02}, \gamma_{03}$) indicate whether specific respondents (e.g., with high θ^{ERS}) take longer and cross-level interaction effects ($\gamma_{11}, \gamma_{21}, \gamma_{31}$) indicate whether specific responses (e.g., $X_{in}^{Extreme}$) take longer for certain levels of latent response style traits (e.g., for high θ^{ERS}).

All analyses were conducted using R (R Core Team, 2019) with Mplus Automation (Hallquist & Wiley, 2018) using Mplus version 7.4 (Muthén & Muthén, 2012) for model fit[3]. Mplus code for model fit is provided on OSF. We set the level of significance to $\alpha = .05$.

Results

Figure 1 and Table 4 provide the estimates of the multilevel analysis for the three datasets. Since response times were log-transformed, the exponential of the estimate (x) is interpreted as a proportional change ($x \times 100\%$) in the dependent variable (see e.g., Lo & Andrews, 2015).

On Level 1, agree and mid responses significantly increased response times, while there is a null effect for extreme responses. Substantively, giving an agree response X_{in}^{Agree} increased response times compared to the respondent's average response time by 28% in Study 1, by 54% in Study 2, and by 21% in Study 3.

[3]Furthermore, we used the packages splitstackshape, gtools, stringr, dplyr, and tidyr for data management (Mahto, 2018; Warnes, Bolker, & Lumley, 2018; Wickham, 2018; Wickham, François, Henry, & Müller, 2018; Wickham & Henry, 2018) as well as gridExtra and ggplot2 for plotting (Auguie, 2017; Wickham, 2016)

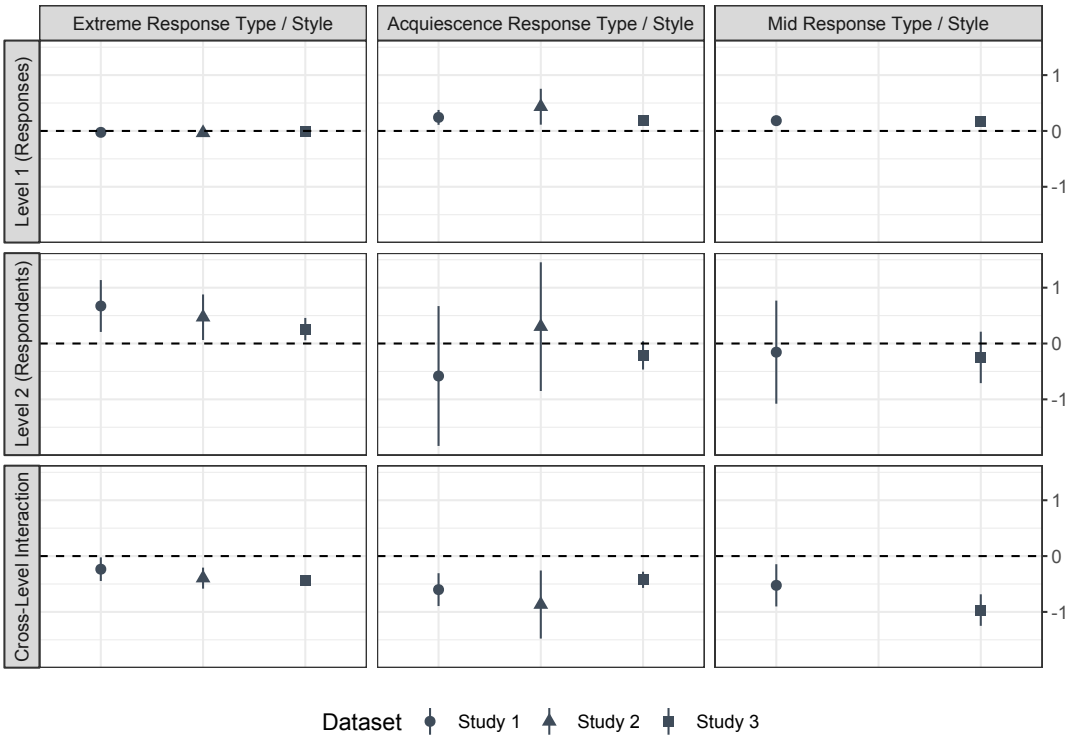

FIGURE 1: Fixed effects estimates of the multilevel analysis (error bars reflect 95% confidence intervals).

Similarly, giving a mid response X_{in}^{Mid} increased the average response time of the respondent by 20% in Study 1, and by 19% in Study 3 compared to a directed response.

On Level 2, there was a significant positive main effect of respondents' ERS levels θ_n^{ERS} on response times in all three datasets. When ERS levels increased by, for example, 0.3 response times increased by 22% in Study 1, by 15% in Study 2, and by 8% in Study 3. There were no significant Level 2 main effects for ARS and MRS, so higher levels of acquiescence or mid response styles did neither increase nor decrease response times.

In all three studies, there was a negative cross-level interaction effect between the type of item responses ($X_{in}^{Extreme}, X_{in}^{Agree}, X_{in}^{Mid}$) and respondents' response styles ($\theta_n^{ERS}, \theta_n^{ARS}, \theta_n^{MRS}$). High levels of response styles in combination with a response that matches the response styles significantly accelerated the response time of the respondent. So, when ERS levels increased by 0.3, and an extreme response was given, respondents were 7%, 11%, or 12% faster in Study 1, 2, and 3, respectively. In case of ARS, an increase of 0.3 in ARS levels jointly with an agree response decreased respondents' response time by 16%, 23%, or 12% in the three datasets. For MRS, a mid response in combination with an increase in MRS

TABLE 4: Summary of Multilevel Model Estimates Predicting Log Response Times

	Study 1			Study 2			Study 3		
Predictors	*B*	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
Level 1 (Responses)									
Intercept (γ_{00})	1.61	0.34	$< .001$	1.38	0.37	$< .001$	1.81	0.11	$< .001$
ERS (γ_{10})	-0.03	0.03	.432	-0.03	0.03	.181	< -0.01	0.02	.803
ARS (γ_{20})	0.24	0.07	$< .001$	0.43	0.16	.008	0.19	0.04	$< .001$
MRS (γ_{30})	0.18	0.05	$< .001$	-	-	-	0.17	0.03	$< .001$
Level 2 (Respondents)									
ERS (γ_{01})	0.67	0.24	.005	0.47	0.21	.024	0.26	0.10	.012
ARS (γ_{02})	-0.58	0.64	.363	0.30	0.59	.607	-0.21	0.13	.100
MRS (γ_{03})	-0.16	0.47	.742	-	-	-	-0.25	0.24	.290
Cross-Level Interaction									
ERS (γ_{11})	-0.24	0.11	.030	-0.40	0.10	$< .001$	-0.44	0.05	$< .001$
ARS (γ_{12})	-0.60	0.15	$< .001$	-0.87	0.31	.005	-0.43	0.07	$< .001$
MRS (γ_{13})	-0.52	0.19	.007	-	-	-	-0.97	0.14	$< .001$
Variance Components									
Intercept (u_{on})	0.09	0.01	$< .001$	0.08	0.01	$< .001$	0.10	0.01	$< .001$
ERS slope (u_{1n})	< 0.01	< 0.01	.771	< 0.01	< 0.01	.914	0.01	< 0.01	.001
ARS slope (u_{2n})	< 0.01	< 0.01	.753	< 0.01	< 0.01	.251	0.01	0.01	.015
MRS slope (u_{3n})	< 0.01	< 0.01	.659	-	-	-	0.02	0.01	.004
Residual (e_{in})	0.11	0.01	$< .001$	0.12	0.01	$< .001$	0.22	0.01	$< .001$

Note. All significance tests are two-sided.

levels by 0.3 decreased response times by 15% in Study 1, and 25% in Study 3. The interpretation of these cross-level interactions will be further illuminated in the following paragraph (see also Figure A1 in Appendix A).

Interpreting Interaction Effects with the Johnson-Neyman Technique

The upper panels of Figures 2, 3, and 4 show raw data scatterplots of response times in seconds (minimum inner 80% quantile) and model-based prediction lines as a function of the latent response style aggregate for extreme, agree, and midpoint responding, respectively. Please note that prediction lines are slightly bent due to reconversion of log response times (that are the basis of the linear model) into response times in seconds. In the lower panel, Johnson-Neyman plots illustrate the change in the effect of an item response on response times as a function of the latent response style aggregate (see Bauer & Curran, 2005; Preacher, Curran, & Bauer, 2006, for details on this technique in multilevel models). For example, the Johnson-Neyman technique displays how the effect of giving an extreme response (X_{in}^{ERS}) on response times (y-axis) changes for different levels of ERS (θ_n^{ERS}; x-axis) and identifies regions of significance, hence regions where the effect is significantly

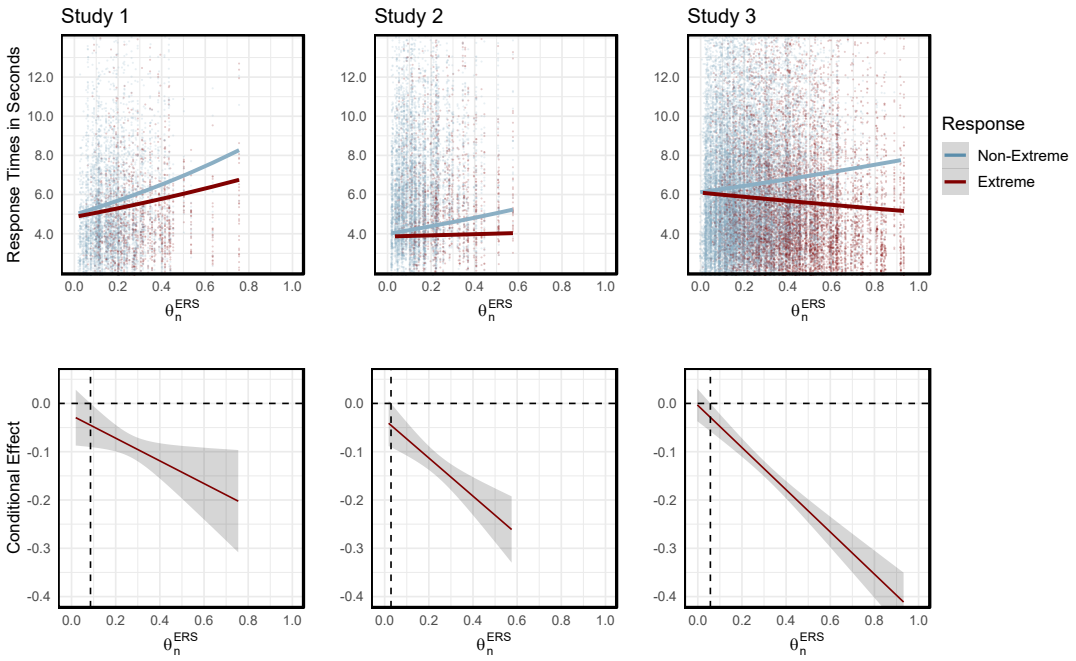

FIGURE 2: Scatterplots with model-based prediction lines (upper panel) and Johnson-Neyman plots (lower panel) to illustrate the effect of Extreme Response Style (ERS) levels and an extreme response on response times.

positive, significantly negative, or not significantly different from zero. Confidence bands represent the uncertainty in the conditional effect and dashed, vertical lines represent the boundaries of the regions of significance.

Extreme Response Style

In the upper panel of Figure 2, the positive Level 2 main effect of θ_n^{ERS} is apparent when averaging over extreme and non-extreme responses. The cross-level interaction leads to the fact that the lines for extreme and non-extreme responses are not parallel. This cross-level interaction is further illustrated in the lower panel using the Johnson-Neyman technique. These plots show the effect of giving an extreme response on response time as a function of the latent ERS estimate on the x-axis. These plots indicate that the higher the ERS level, the stronger was the negative effect of extreme compared to non-extreme responses on response times. This conditional effect was significantly negative for $\theta_n^{ERS} > .09$ across datasets as illustrated by the dashed line marking the boundary of the region of significance. Very low levels of ERS do not impact the effect of an extreme response on response time. Stated differently, responses were slowest when respondents with high ERS levels selected a non-extreme response category, which seems to be a more carefully considered category choice the higher the ERS level.

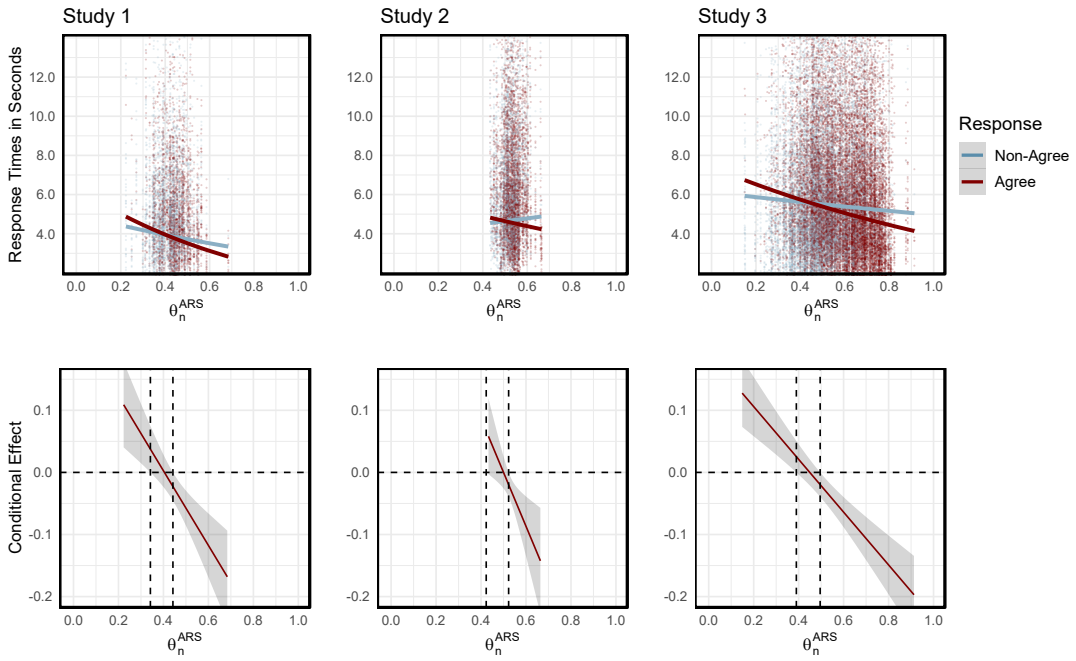

FIGURE 3: Scatterplots with model-based prediction lines (upper panel) and Johnson-Neyman plots (lower panel) to illustrate the effect of Acquiescence Response Style (ARS) levels and an agree response on response times.

Acquiescence Response Style

Figure 3 illustrates the interaction effect for ARS which followed a disordinal pattern. Hence, for low ARS levels giving an agree response increased response times, while for high ARS levels giving an agree response decreased response times. Across studies, the conditional effect was significantly positive for $\theta_n^{ARS} < .34$ and significantly negative for $\theta_n^{ARS} > .52$. Hence, responses were faster when respondents with low ARS selected a non-agree response category and respondents with high ARS levels selected an agree response category.

Mid Response Style

Figure 4 shows the interaction effect for MRS in the studies with an odd number of response categories. In Study 1, the effect of MRS responses on the effect of MRS latent aggregate on response time was significant for low levels of MRS ($\theta_n^{MRS} < .29$), where response times increased when a mid response was given. The upper boundary was $\theta_n^{MRS} > .69$, implying that for MRS levels above this boundary response times decreased when a mid response was given even though there was no data available for this range of MRS in the dataset. We see a pronounced disordinal interaction in Study 3 indicating that giving a midpoint response increased response times for low MRS levels ($\theta_n^{MRS} < .15$), while it

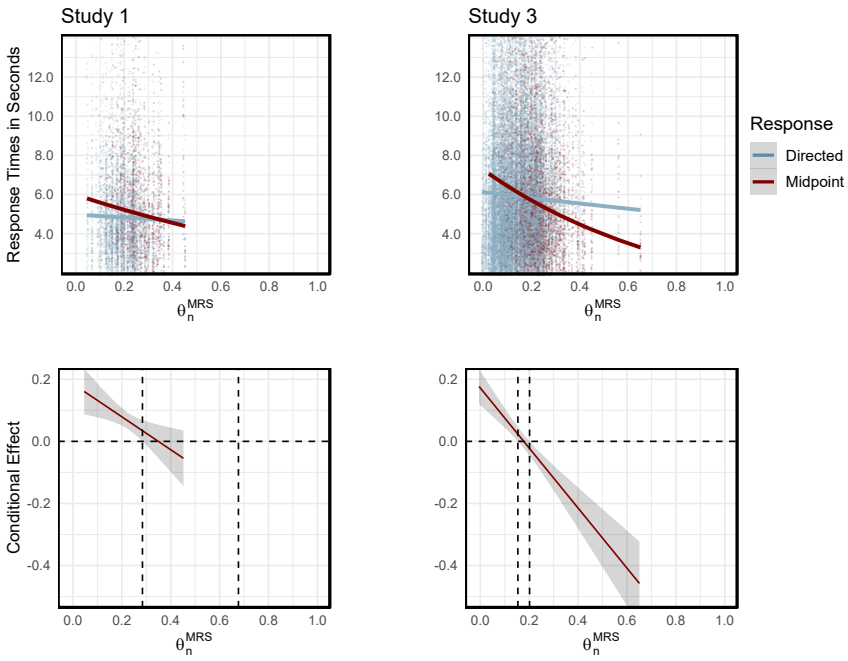

FIGURE 4: Scatterplots with model-based prediction lines (upper panel) and Johnson-Neyman plots (lower panel) to illustrate the effect of Mid Response Style (MRS) levels and a mid response on response times (no MRS effect was modeled in Study 2 due to the use of a rating scale with an even number of categories).

increased response times for MRS levels higher than $\theta_n^{MRS} > .20$. We can conclude that lower MRS levels lead to higher response times when a midpoint response was given, while a mid response for higher MRS trait levels results in shorter response times.

Discussion

In this research project, we investigated the effects of extreme, agree, and mid responding on response times. Although response times are frequently used to describe cognitive processes, they have rarely been linked to personality traits or response tendencies. However, response times can provide useful insights into the cognitive processes underlying rating scale usage and the use of response styles. We employed a multilevel modeling approach and investigated the effect of item responses, respondents' response styles and their cross-level interaction on response times in three different studies.

Results Interpretation

On the level of individual item responses, we investigated the effect of dichotomous indicators of extreme, agree, and mid responses on response times and found consistent main effects across the three studies. In contrast to Hypothesis 1a, there was no difference in response times between extreme and non-extreme responses in any of the datasets which contradicts the results by Casey and Tryon (2001). In accordance with Hypothesis 1b, response times increased when agree responses were given. This is in line with evidence presented by Swain et al. (2008), Hanley (1965), and Rogers (1973) indicating that agree responses might be related to cognitive burden. Similarly, response times increased when a midpoint response was given which is in line with Hypothesis 1c. Hence, choosing the midpoint seems to be a deliberate process where respondent weigh the different alternatives, and choose the midpoint as a final response. The results corroborate findings by Kulas and Stachowski (2009) indicating that the midpoint was the response option with the longest response latency.

On the level of the respondent, we explored the influence of response style traits ERS, ARS, and MRS on response times (Exploratory Analyses 2a-c). In all three datasets, we found a positive main effect of the ERS trait on response times. Thus, the higher the trait, the more time does the respondent take to respond. Particularly when responses are non-extreme, respondents with high ERS levels seem to take more time to respond. No main effects were found for ARS in any of the three datasets contradicting the results by Mayerl (2013) and the descriptive results by Knowles and Condon (1999). Neither did we find a main effect for MRS on the respondent level.

The multilevel analysis used here yielded original evidence for cross-level interactions, hence matching effects of response styles and item responses. As predicted in Hypotheses 3a-c, there were significant negative cross-level interaction effects of item responses and response style traits on response times across all datasets and across all response styles. Thus, giving a response that is in line with the response style trait decreases response times or, stated differently, the response style trait facilitates the choice of certain categories in terms of response speed. The illustration with the Johnson-Neyman technique (Figures 2, 3, and 4) also brought novel insights into the range of response style trait levels ($\theta_n^{ERS}, \theta_n^{ARS}, \theta_n^{MRS}$) for which category choices were affected or unaffected. Please note that when respondents' latent response style trait lies in the area over which the cross-level interaction effect is not significantly different from zero (area within the boundaries of the region of significance), response times are equal for both response options (e.g.,

an extreme or a non-extreme response). This area might therefore demarcate the range over which response styles have the smallest impact on the response. In our analyses, these "neutral" response style levels were identified to be very low for ERS across datasets ($\theta_n^{ERS} < .09$), indicated by an ordinal interaction effect. In contrast, for ARS and MRS moderate response style trait levels were identified as neutral (ARS: $.34 < \theta_n^{ARS} < .52$; MRS: $.15 < \theta_n^{MRS} < .20$; MRS in Study 3), as indicated by a disordinal interaction effect. Across all three response styles, the range of these neutral levels was very small. Therefore, a preference (avoidance) for certain response category types is consequential for a majority of respondents and there exists almost no level of response style for which the category choice is not facilitated by response tendencies (see also Figure A1 in Appendix A for an illustration of the frequency of different effects of response types on response times in the datasets). The small range of response style levels indicates that for nearly all respondents, response styles facilitate certain category choices in terms of response speed.

Theoretical Implications

Cognitive processes underlying response style usage

The analyses and results of the current investigation show that extreme responding is qualitatively distinct from acquiescent and mid responding and follows a different cognitive process. Based on the visualization of the Johnson-Neyman technique, respondents with moderate and high ERS trait levels take longer to give non-extreme responses (see Figure 2). Furthermore, only at very low ERS trait levels, extreme and non-extreme responses have similar response times. Since overall high ERS trait levels are accompanied by longer response times, the results do not support the notion that extreme response style is associated with low cognitive effort of the respondent. In contrast, the positive main effect and negative cross-level interaction rather indicate that respondents with moderate to high ERS levels give non-extreme responses more deliberately.

In contrast, acquiescent and mid responding show very similar patterns of response processes. First, on Level 1, we found positive main effects of agree and mid responses, indicating these responses go along with longer response times. Second, there were no main effects of the ARS and MRS traits, indicating that across responses, differences in respondents' ARS and MRS levels did not explain differences in response times. Third, disordinal interactions were found for acquiescent and mid responding indicating that responses that are in line with the

respective response style are faster than responses that contradict the response style.

Knowles and Condon (1999) reported that response times were faster when ARS-respondents agreed with the item and based their argumentation on a dual process theory of acquiescence. According to this theory, people either agree with the item instantly without investing any effort, or follow a normal processing route including comprehension, reconsideration and decision phases that require more time and effort. We were able to replicate the finding by Knowles and Condon for high ARS levels. At the same time, our data showed a similar pattern for respondents with low ARS levels who were faster when they disagreed (see Figure 3 and A1 in Appendix A). This is a clear contradiction to a dual process theory with a unipolar conceptualization of acquiescence where the absence of acquiescence means moderate responding (Knowles & Condon, 1999, see also Plieninger & Heck, 2018). The results rather suggest a bipolar acquiescence construct where respondents with low levels of acquiescence tend to disagree with items more easily, while respondents with high levels of acquiescence tend to agree with items independent of item content. The same process seems to hold for MRS: the disordinal cross-level interaction for MRS indicates that mid responses are slower for low MRS levels, and may be faster for high MRS levels compared to directed responses. Hence, we replicated the effect that low MRS trait levels lead to higher response times when giving a midpoint response (Kulas & Stachowski, 2009) and extended this effect by differentiating between areas of significance for different MRS levels (see Figure 4 and A1 in Appendix A). Please note that the variance of the latent MRS aggregate was smaller than for the other response styles; a phenomenon that is commonly observed in response style measurement (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Wang, Wilson, & Shih, 2006).

Speed-distance hypothesis

The analyses of response styles and response times have shown that not only personality traits, but also response styles follow the speed-distance hypothesis: the more likely a response is for a certain respondent, the faster he or she gives this type of response. The results suggest that the effect on response times is due to a higher confidence in the response when respondents follow their response tendency (i.e. self-schemata) which guides responses and decreases response times (McIntyre, 2011). In contrast, giving a response that is contrary to the respondent's response style level increases difficulty and therefore leads to longer response times (Dunn et al., 1972; Kuiper, 1981). The speed-distance hypothesis is a robust theory

with precise predictions in many fields besides personality research, for example in signal detection theory (Maddox, Ashby, & Gottlob, 1998) or value research (Bilsky, Borg, Janik, & Groenen, 2013). With the present investigation, we further extend the application of the speed-distance hypothesis and present evidence for its validity in the area of response styles.

Strengths and Limitations of the Current Analysis

The strength of this study is the comparison of three different datasets that consist of different item types, samples, and response category numbers. The fact that results are highly consistent across the three datasets is even more remarkable given the differences between the data sources. Study 1 only used heterogeneous items that refer to different content domains and therefore is ideal to measure response styles (De Beuckelaer et al., 2010; Greenleaf, 1992). However, the sample size with $N = 161$ respondents and $I = 39$ items is sufficient, but not abundant. Since Study 2 combined items of two personality scales with heterogeneous items, it is well suited to measure response styles, while at the same time being generalizable to applied settings on the basis of the two personality scales. However, as a 6-category scale was used, MRS cannot be measured in this study. Study 3 contained homogeneous items assessing five different traits from organizational psychology. Since intercorrelations between items were moderate (mean absolute correlation: $r = .23$ in contrast to $r = .11$ in Study 1 and 2), response styles can be measured across the different content scales (see also Wetzel & Carstensen, 2017, for a discussion on response style measurement across scales). Study 3 demonstrates that the results obtained in Study 1 and Study 2 are generalizable to applied measurement contexts. Besides the applied context in which the study was conducted, the main advantages of this dataset is the large sample size, the high variances of response style traits, and a large power. Overall, the high consistency of effects between these different data sources underpins the results' robustness, stability, and generalizability.

The positive main effect that respondents with high ERS levels take more time to respond is a result of an exploratory analysis and contradicts previous assumptions and findings in the literature (Aichholzer, 2013; Casey & Tryon, 2001; Krosnick, 1999). The result suggests that high ERS levels may be associated with an increased rather than decreased cognitive effort, but more studies are necessary to further test and corroborate this effect.

A major challenge when analyzing response times is the noise that is inherent in the data (Fazio, 1990; Lo & Andrews, 2015; Ratcliff, 1993). With our multilevel

modeling approach, we were able to separate variance components in response times that are due to specific responses (Level 1), respondents' response style traits (Level 2) and their cross-level interactions. Before the main analyses, we made several choices to preprocess response time data, such as excluding responses to initially omitted items when respondents where redirected to the survey page, responses correcting previously given responses, and response time outliers[4]. Across all preprocessing steps, we paid careful attention to use procedures that are well embedded in the response times literature connected to rating scale responses (Bassili & Fletcher, 1991; Höhne & Schlosser, 2018; Mayerl & Urban, 2008; Mulligan et al., 2003), and applied the same procedures in all three studies.

In this analysis, response times served as indicators of response processes, for example of spontaneous or deliberate response modes. However, response times are not pure process measures. When interpreting changes in response times, one must be aware that implications are based on assumptions on the relation of response times and cognitive processes. The relation of response times and cognitive processes are substantiated by evidence in the literature (see e.g., Lo & Andrews, 2015), but remain presumed associations as processes themselves are always unobserved.

Directions for Future Research

This research project opens up new areas for future research. While we focused on extreme, acquiescent, and mid responding as response tendencies that occur in rating scale measurement, other response biases such as social desirable or careless responding (Andersen & Mayerl, 2017; Dunn et al., 1972; Ellingson, Smith, & Sacket, 2001; Meade & Craig, 2012) may similarly be analyzed with respect to response times. Besides, other process measures, such as eye-tracking and mouse-tracking or even fMRI and EEG measures, could provide useful insights into cognitive processes in rating scale usage. As response times, these process measurement methods differentiate between spontaneous and deliberate response processes, but may also provide information on the guidance of attention, such as whether respondents reread a question, or encounter difficulties in the response mapping process (Franco-Watkins & Johnson, 2011; Kamoen, Holleman, Mak,

[4]As a robustness check, we reanalyzed the three samples after excluding responses to items that received more than one click (rather than keeping the first, spontaneous response). The pattern of the estimates remained unchanged, but two effects were no longer significant. This is attributable to losing power when the sample size is reduced, which is corroborated by the fact that significance was not affected in Study 3, which had the largest power. The effects which were no longer significant were the interaction effect for ERS in Study 1 and the Level 1 effect of agree responses in Study 2.

Sanders, & van den Bergh, 2011; van Hooft & Born, 2012). fMRI and EEG measures may additionally provide insights into physiological correlates of response speed.

The relation of response styles and response times might inform the measurement of content traits and of response biases. So far, response times inform the measurement of personality traits and increase, for example, test information (Ferrando & Lorenzo-Seva, 2007; Ranger & Ortner, 2011). However, our results show that response times are not only indicators of the cognitive process with regard to the content trait, but also with regard to response styles. Thus, response times measure several processes: response processes related to the content of the items as well as processes underlying response style usage. Future research should evaluate the potential to improve measurement of personality variables when incorporating response style as well as response time information.

While our main goal was to describe how response styles manifest themselves in the response process, the findings presented in this article may lead to further investigations providing practical guidance for applied measurement situations. With this regard, response times may be analyzed using data originating from experiments in which certain assessment characteristics are manipulated. For instance, one could vary the number of response categories or present items in random order to assess differences in response style effects on response times in different measurement settings.

Furthermore, we did not only collect response times for each item, but also for each mouse click that the respondent made on the survey page. Thus, collecting response times in such a way may provide useful information in test construction and item selection. Response time data of this kind allows one to evaluate whether responses to specific items were changed more often, or whether reversed-coded items are difficult to process cognitively. Furthermore, changing a given response may be an indicator of high deliberation and motivation of the respondent with regards to the survey. Hence, future research could evaluate whether such correction of responses may be negatively related to careless responding (Meade & Craig, 2012).

Conclusion

Our analyses have shown that agree and midpoint responding follow a joint cognitive process that is qualitatively different from extreme responding: respondents need more time to give agree and midpoint, but not extreme, responses and respondents with high ERS traits take more time to respond, while this is not

the case for respondents with high ARS and MRS traits. However, extreme, acquiescent, and midpoint response styles accelerate response times when the given response is in line with the latent response style trait. This finding indicates that when respondents follow their response styles, their self-schemata guide and therewith accelerate item responses as proposed by the speed-distance hypothesis. Our analyses suggest that every respondent employs some type of response tendency when reacting to a rating scale and that the area of a neutral response is actually quite small. The joint result of our studies may furthermore guide future developments in designing testing situations to improve psychological assessment.

References

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

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*, 957–970. doi:10.1016/j.ssresearch.2013.01.002

Akrami, N., Hedlund, L. E., & Ekehammar, B. (2007). Personality scale response latencies as self-schema indicators: The inverted-U effect revisited. *Personality and Individual Differences*, *43*, 611–618. doi:10.1016/j.paid.2006.12.005

Andersen, H., & Mayerl, J. (2017). Social desirability and undesirability effects on survey response latencies. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *135*, 68–89. doi:10.1177/0759106317710858

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics. *R package version 2.3*. Retrieved from https://cran.r-project.org/package=gridExtra

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

Bassili, J. N., & Fletcher, J. F. (1991). Response-time measurement in survey research: A new method for Cati and a new look at nonattitudes. *Public Opinion Quarterly*, *55*(3), 331–346.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*, 390–399. doi:10.1086/297760

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400. doi:10.1207/s15327906mbr4003_5

Bilsky, W., Borg, I., Janik, M., & Groenen, P. (2013). Children's value structures - Imposing theory-based regional restrictions onto an ordinal MDS solution.

In *14th facet theory conference* (pp. 25–40). Recife, Brazil: Proceedings of the 14th Facet Theory Conference.

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

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

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

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

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

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

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin de Méthodologie Sociologique*, *103*, 5–25. doi:10.1177/075910630910300103

Casey, M. M., & Tryon, W. W. (2001). Validating a double-press method for computer administration of personality inventory items. *Psychological Assessment*, *13*, 521–530. doi:10.1037/1040-3590.13.4.521

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. doi:10.1037/h0040372

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity*, *44*, 761–775. doi:10.1007/s11135-009-9225-z

Dunn, T. G., Lushene, R. E., & O'Neil, H. F. (1972). Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology*, *39*, 381–387. doi:10.1037/h0033855

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

Eisenberger, R., Huntington, R., Hutchison, S., & Sowa, D. (1986). Perceived organizational support. *Journal of Applied Psychology*, *71*, 500–507. doi:10.1037//0021-9010.71.3.500

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

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Research Methods in Personality and Social Psychology* (pp. 74–97). SAGE Publications.

Fekken, G. C., & Holden, R. R. (1992). Response latency evidence for viewing personality traits as schema indicators. *Journal of Research in Personality*, *26*, 103–120. doi:10.1016/0092-6566(92)90047-8

Fekken, G. C., & Holden, R. R. (1994). The construct validity of differential response latencies in structured personality tests. *Canadian Journal of Behavioural Science*, *26*, 104–120. doi:10.1037/0008-400X.26.1.104

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item-response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543. doi:10.1177/0146621606295197

Fladerer, M., & Misterek, L. (2018). Identity leadership and burnout: A multilevel mediation study. Manuscript in preparation. Retrieved from osf.io/4x9qg

Franco-Watkins, A. M., & Johnson, J. G. (2011). Decision moving window: Using interactive eye tracking to examine decision processes. *Behavior Research Methods*, *43*, 853–63. doi:10.3758/s13428-011-0083-y

Fuhrman, R. W., & Funder, D. C. (1995). Convergence between self and peer in the response-time processing of trait-relevant information. *Journal of Personality and Social Psychology*, *69*, 961–974. doi:10.1037/0022-3514.69.5.961

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*, 349–360. doi:10.1093/poq/nfp031

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly*, *56*, 328–351. doi:10.1086/269326

Grimm, S., & Church, A. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality*, *33*, 415–441. doi:10.1006/jrpe.1999.2256

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, 1–18. doi:10.1080/10705511.2017.1402334

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

Hanley, C. (1965). Personality item difficulty and acquiescence. *Journal of Applied Psychology*, *49*, 205–208. doi:10.1037/h0022107

He, J., & Van De Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Difference*, *55*, 794–800. doi:10.1016/j.paid.2013.06.017

Heck, D. W., & Erdfelder, E. (2016). Extending multinomial processing tree models to measure the relative speed of cognitive processes. *Psychonomic Bulletin & Review*, *23*, 1440–1465. doi:10.3758/s13423-016-1025-6

Hibbing, M. V., Cawvey, M., Deol, R., Bloeser, A. J., & Mondak, J. J. (2017). The relationship between personality and response patterns on public opinion surveys: The Big Five, extreme response style, and acquiescence response style. *International Journal of Public Opinion Research*. doi:10.1093/ijpor/edx005

Höhne, J. K., & Schlosser, S. (2018). Investigating the adequacy of response time outlier definitions in computer-based web surveys using paradata SurveyFocus. *Social Science Computer Review*, *36*, 369–378. doi:10.1177/0894439317710450

Kamoen, N., Holleman, B., Mak, P., Sanders, T., & van den Bergh, H. (2011). Agree or disagree? Cognitive processes in answering contrastive survey questions. *Discourse Processes*, *48*, 355–385. doi:10.1080/0163853X.2011.578910

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Jornal of Personality and Social Psychology*, *77*, 379–386. doi:10.1037/0022-3514.77.2.379

Krosnick, J. A. (1999). Survey Research. *Annual Review of Psychology*, *50*, 537–567. doi:10.1146/annurev.psych.50.1.537

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype: The inverted U RT effect for self and other judgments. *Personality and Social Psychology Bulletin*, *7*, 438–443. doi:10.1177/014616728173012

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. *Journal of Research in Personality*, *43*, 489–493. doi:10.1016/j.jrp.2008.12.005

Lee, K., & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment*, *18*, 182–191. doi:10.1037/1040-3590.18.2.182

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*, 1003–1020. doi:10.1002/acp.1602

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1–16. doi:10.3389/fpsyg.2015.01171

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*, 444–467. doi:10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. doi:10.1037/a0012869

Machunsky, M., & Meiser, T. (2006). Personal Need for Structure als differenzialpsychologisches Konstrukt in der Sozialpsychologie. *Zeitschrift für Sozialpsychologie*, *37*(2), 87–97. doi:10.1024/0044-3514.37.2.87

Maddox, W. T., Ashby, F. G., & Gottlob, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception and Psychophysics*, *60*, 620–637. doi:10.3758/BF03206050

Mael, F., & Asiforth, B. E. (1992). Alumni and their alma mater: A partial test of the reformulated model of organizational identifiation. *Journal of Organizational Behavior*, *13*, 103–123. doi:10.1002/job.4030130202

Mahto, A. (2018). splitstackshape: Stack and reshape datasets after splitting concatenated values. *R package version 1.4.6*. Retrieved from https://cran.r-project.org/package=splitstackshape

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation appraoches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764–802. doi:10.1080/00273170903333665

Mayerl, J. (2013). Response latency measurement in surveys. Detecting strong attitudes and response effects. *Survey Methods: Insights from the Field*, 1–27. doi:10.13094/SMIF-2013-00005

Mayerl, J., & Urban, D. (2008). *Antwortreaktionszeiten in Survey-Analysen* (1. Auflage). Wiesbaden: VS Verlag für Sozialwissenschaften.

McIntyre, H. H. (2011). Investigating response styles in self-report personality data via a joint structural equation mixture modeling of item responses and response times. *Personality and Individual Differences*, *50*, 597–602. doi:10.1016/j.paid.2010.12.001

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. doi:10.1037/a0028085

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*, 1539–1550. doi:10.1016/j.paid.2008.01.010

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

Moors, G. (2008). Exploring the effect of a middle response category on response style in attitude measurement. *Quality and Quantity*, *42*(6), 779–794. doi:10.1007/s11135-006-9067-x

Mulligan, K., Grant, J. T., Mockabee, S. T., & Monson, J. Q. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, *11*, 289–301. doi:10.1093/pan/mpg004

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

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality predictors of extreme response style. *Journal of Personality*, *77*, 261–286. doi:10.1111/j.1467-6494.2008.00545.x

Neubauer, A. C., & Malle, B. F. (1997). Questionnaire response latencies: Implications for personality assessment and self-schema theory. *European Journal of Psychological Assessment*, *13*, 109–117. doi:10.1027/1015-5759.13.2.109

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). doi:10.1016/B978-0-12-590241-0.50006-X

Pfister, M. (2018). Are extreme and acquiescent response styles related to the implicit self-concept of personality? Unpublished Bachelor Thesis. Mannheim, Germany: University of Mannheim.

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

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

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448. doi:10.3102/10769986031004437

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

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*, 389–406. doi:10.1177/0013164410382895

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*, 510–532. doi:10.1037/0033-2909.114.3.510

Riggs, M. L., Warka, J., Babasa, B., Betancourt, R., & S., H. (1994). Development and validation of self-efficacy and outcome expectancy scales for job-related applications. *Educational and Psychological Measurement*, *54*, 793–802. doi:10.1177/0013164494054003026

Rogers, T. B. (1973). Toward a definition of the difficulty of a personality item. *Psychological Reports*, *33*, 159–166. doi:10.2466/pr0.1973.33.1.159

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

Sarubin, N., Gutt, D., Giegling, I., Böhner, M., Hilbert, S., Krähenmann, O., ... Padberg, F. (2015). Erste Analyse der psychometrischen Eigenschaften und Struktur der deutschsprachigen 10-und 25-Item Version der Connor-Davidson Resilience Scale (CD-RISC). *Zeitschrift für Gesundheitspsychologie*, *23*(3), 112–122. doi:10.1026/0943-8149/a000142

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The application of factorial surveys in general population samples: The effects of respondent age and education on response times and response consistency. *Survey Research Methods*, *5*, 89–102. doi:10.18148/srm/2011.v5i3.4625

Steffens, N. K., Haslam, S. A., Reicher, S. D., Platow, M. J., Fransen, K., Yang, J., ... Boen, F. (2014). Leadership as social identity management: Introducing the Identity Leadership Inventory (ILI) to assess and validate a four-dimensional model. *Leadership Quarterly*, *25*, 1001–1024. doi:10.1016/j.leaqua.2014.05.002

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*, 116–131. doi:10.1509/jmkr.45.1.116

Tourangeau, R., Rasinski, K. A., & D'Andrade, R. (1991). Attitude structure and belief accessibility. *Journal of Experimental Social Psychology*, *27*, 48–75. doi:10.1016/0022-1031(91)90010-4

van der Linden, W. J., Klein Entink, R. H., & Fox, J.-P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. doi:10.1177/0146621609349800

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

van Dijk, T. K., Datema, F., Piggen, A.-L. J. H. F., Welten, S. C. M., & van de Vijver, F. J. R. (2009). Acquiescence and extremity in cross-national surveys: domain dependence and country-level correlates. In A. Gari & K. Mylonas (Eds.), *Quod erat demonstrandum: From herodotus' ethnographic journeys to cross-cultural research: Proceedings from the 19th international congress of the international association for cross-cultural psychology* (pp. 149–158). Athens, Greece: Pedio Books. Retrieved from https://scholarworks.gvsu.edu/iaccp_papers/51/

van Hooft, E. A., & Born, M. P. (2012). Intentional response distortion on personality tests: Using eye-tracking to understand response processes when faking. *Journal of Applied Psychology*, *97*, 301–316. doi:10.1037/a0025711

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

Warnes, G. R., Bolker, B., & Lumley, T. (2018). gtools: Various R programming tools. *R package version 3.8.1*. Retrieved from https://cran.r-project.org/package=gtools

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response

category labels. *International Journal of Research in Marketing*, *27*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

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., & 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

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2018). stringr: Simple, consistent wrappers for common string operations. *R package version 1.3.1*. Retrieved from https://cran.r-project.org/package=stringr

Wickham, H., François, R., Henry, L., & Müller, K. (2018). dplyr: A grammar of data manipulation. *R package version 0.7.6*. Retrieved from https://cran.r-project.org/package=dplyr

Wickham, H., & Henry, L. (2018). tidyr: Easily tidy data with 'spread()' and 'gather()' functions. *R package version 0.8.1*. Retrieved from https://cran.r-project.org/package=tidyr

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*, 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38. doi:10.1111/j.1745-3984.2006.00002.x

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *21*, 51–68. doi:10.1002/acp.1331

Zhang, C., & Conrad, F. G. (2013). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*, 127–135. doi:10.18148/srm/2014.v8i2.5453

Appendix A

Magnitude of the Cross-Level Interaction Effect on Response Times

Figure A1 illustrates the predicted magnitude of the effect of response type (extreme, agree, mid responses) on response times in the three datasets. On the x-axis, we see the percentage change in response times when the respondent gives an extreme response compared to a non-extreme response (upper row), an agree response compared to a non-agree response (middle row) or a mid response compared to a directed response (lower row), hence given a certain response style trait level. When this effect is negative, response times decrease when the respondents gives a certain response (extreme, agree, or midpoint); when it is positive, response times increase when the respondent gives a certain response. On the y-axis, we see the frequency of respondents in the sample for whom this effect takes place. For example, in the upper row the second bar from the right in the Pfister (2018) data indicates that for more than 50 respondents in the sample, response times decreased by approximately 3% when giving an extreme response compared to a non-extreme response. We can see that for extreme responses, the effect is negative for the whole sample (see also Figure 2). In contrast, for agree responses, negative as well as positive effects have occurred (see the disordinal interaction in Figure 3); the same applies to mid responses where response times increased for a majority of the sample, but also decreased for a subsample, when giving a mid response (see the disordinal interaction in Figure 4).

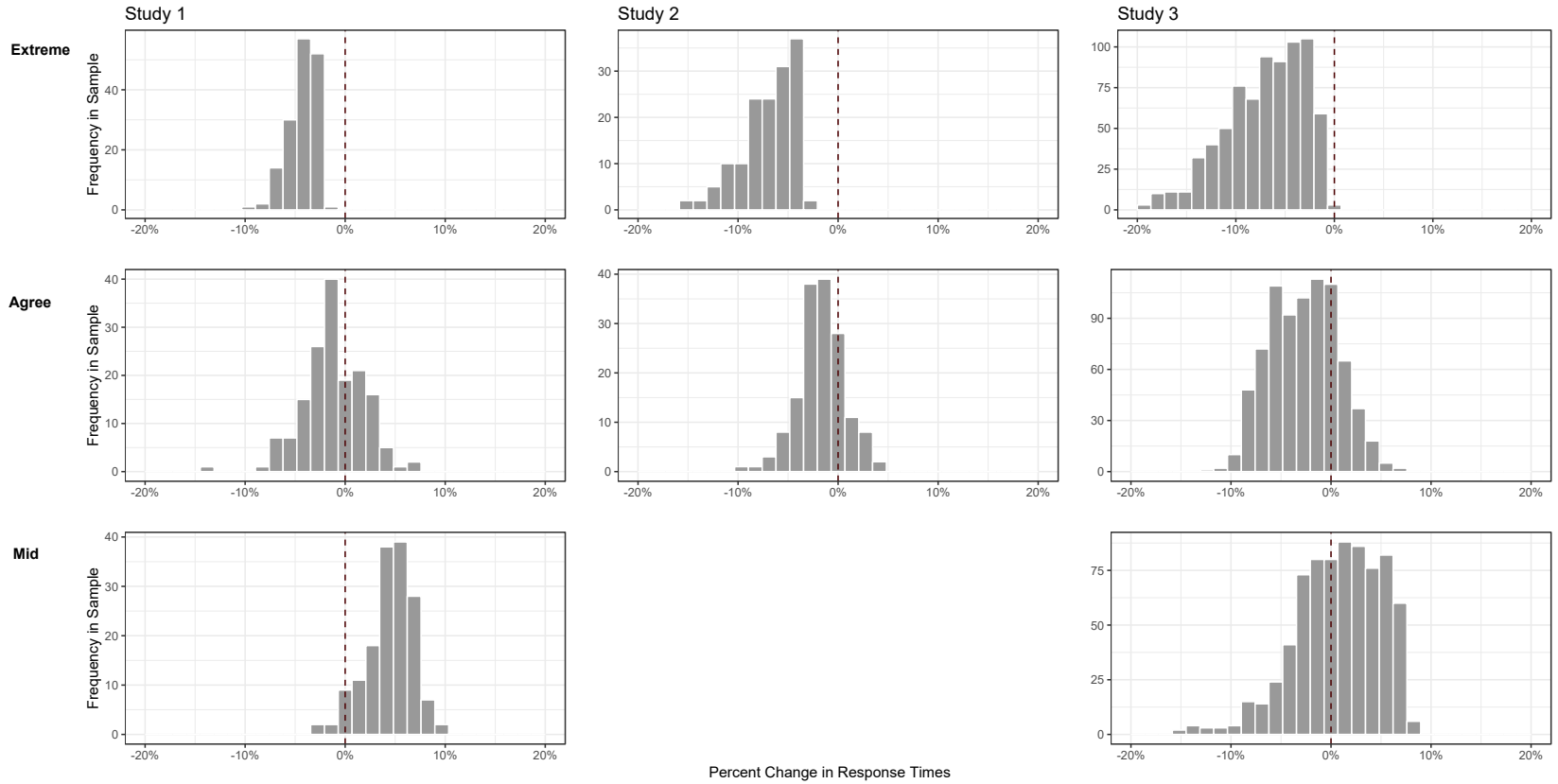

FIGURE A1: Histogram plots illustrating cross-level interaction effects in terms magnitude of the impact of response type (extreme, agree, midpoint) on response times on the respondent level in the three datasets; x-axis shows the percentage change in response times for one respondent, hence given a certain response style trait level, when giving an extreme compared to a non-extreme (upper row), an agree compared to a non-agree (middle row), or a midpoint compared to a directed (lower row) response, y-axis shows the frequency of occurrence of this effect in the analysis sample.