

# Improving the Measurement and Understanding of Response Process Heterogeneity by IRTree Modeling

VIOLA MERHOF

*Inaugural Dissertation*

Submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group “Statistical Modeling in Psychology” at the University of Mannheim

*1<sup>th</sup> Supervisor:*

Prof. Dr. Thorsten Meiser

*2<sup>th</sup> Supervisor:*

Prof. Dr. Beatrice G. Kuhlmann

*Dean of the School of Social Sciences:*

Prof. Dr. Michael Diehl

*Thesis Reviewers:*

Prof. Dr. Eunike Wetzel

Prof. Dr. Florian Keusch

*Thesis Defense:*

May 21, 2024

Für meine Familie



# Contents

<b>Summary</b>	<b>VII</b>
<b>Articles</b>	<b>IX</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 IRTree Modeling</b>	<b>7</b>
2.1 Parameterizations of Response Processes and their Heterogeneity . . . . .	8
2.2 Separation of Response Processes . . . . .	11
<b>3 Modeling Heterogeneity of Response Processes</b>	<b>15</b>
3.1 Dynamic Within-Person Heterogeneity of Response Process Involvement .	16
3.2 Heterogeneity of Item Response Functions . . . . .	19
<b>4 Illustrative Application</b>	<b>25</b>
4.1 Separation of Trait and Extreme Response Style . . . . .	25
4.2 Dynamic Response Process Involvement . . . . .	26
4.3 Dominance and Ideal Point Item Response Functions . . . . .	28
<b>5 General Discussion</b>	<b>31</b>
5.1 Measurement and Understanding of Response Process Heterogeneity by IRTree Modeling . . . . .	31
5.2 Limitations and Future Directions . . . . .	33
5.3 Conclusion . . . . .	35
<b>6 Bibliography</b>	<b>37</b>
<b>A Data Set, Analysis, and Model Specifications of Illustrative Application</b>	<b>45</b>
A.1 Data Set . . . . .	45
A.2 Analysis . . . . .	45
A.3 Models Related to Article I . . . . .	46
A.4 Models Related to Article II . . . . .	47
A.5 Models Related to Article III . . . . .	49
<b>B Copies of Articles</b>	<b>51</b>
<b>C Acknowledgements</b>	<b>159</b>



## Summary

Self-reported rating responses provide valuable information on the characteristics of a person that cannot be observed directly, and numerous psychometric approaches are available to derive individual trait estimates from the given item responses. However, the mapping of latent trait levels and manifest answers may be biased if, in addition to trait-based responding, trait-unrelated response processes such as response styles affect the category selection. This bias may further be exacerbated by additional heterogeneity of the involved response processes, such as when the relevance of latent characteristics for the respondents' judgments differs between measurement units. Heterogeneous response processes may result in distorted trait measurements and misinterpretations of cognitive processes underlying item responding whenever the complexity of the respondents' behavior is not adequately reflected in the analysis model.

In this thesis, I demonstrate how item response tree (IRTree) modeling can be used to address the heterogeneity of response processes. In the first article, I focus on heterogeneity with regard to the person characteristics on which the response processes are based, and I show how IRTree models can be specified to effectively disentangle the influences of traits and response styles from each other. In the second article, I examine systematically changing influences of response processes throughout a questionnaire, and I develop dynamic IRTree models that incorporate such changes. In the third article, I consider that the involved response processes may adhere to heterogeneous item response functions, and I propose a general IRTree framework that can incorporate the combined influences of both dominance and ideal point response processes.

The conducted research highlights the importance of modeling heterogeneous response processes to enhance the measurement of latent characteristics and to foster the interpretability of model parameters. Thereby, the flexibility of the IRTree model class for addressing heterogeneous response processes is illustrated and extended by new developments. Overall, this dissertation contributes to the field of psychometrics by providing a tool to improve the measurement and the understanding of response process heterogeneity, and thus, to increase the validity of assessments through rating scales.





## Articles

This thesis is the result of research conducted in the context of the research training group "Statistical Modeling in Psychology" (SMiP). It is based on three articles, two of which have been published and one has been submitted for publication. Copies of the articles are appended to this dissertation.

### ARTICLE I

Merhof, V., Böhm, C. M., & Meiser, T. (2023). Separation of traits and extreme response style in IRTree models: The role of mimicry effects for the meaningful interpretation of estimates. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/00131644231213319>

### ARTICLE II

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, *88*(4), 1354-1380. <https://doi.org/10.1007/s11336-023-09901-0>

### ARTICLE III

Merhof, V., & Meiser, T. (2023). *Co-occurring dominance and ideal point response processes: A general IRTree framework for multidimensional item responding*. Revision invited by Behavior Research Methods.



# 1 Introduction

*“The numbers may vary, the ratings diverge,  
But the truth lies within, waiting to emerge.”*

— ChatGPT (OpenAI, 2023)

Rating scales are an essential tool in psychology and the social sciences. For more than 300 years,<sup>1</sup> they have been used to assess a variety of person characteristics and are nowadays indispensable in both research and applied fields (Heiser, 2023; McReynolds & Ludwig, 1987). The number of questionnaires measuring opinions, beliefs, preferences, attitudes, and other subjective experiences is arguably difficult to count, but can be considered quite substantial, and most definitely constantly growing. Therefore, it is no surprise that searching Google Scholar for the term *rating scale* yields over 6 million entries and entering the same into Google’s standard search engine even gives over 800 million results (retrieved in November 2023).

The extensive use of rating scales can be attributed to the fact that they provide a straightforward method to quantify a wide range of person-specific traits that cannot be directly observed. Thereby, the fundamental premise is that the true latent trait levels are reflected in the self-reported responses,<sup>2</sup> and consequently, that respondents derive their answers from subjective trait-related information. However, this mapping of substantive traits and rating responses can be systematically biased if, in addition to the trait-based response process, further processes influence the respondents’ category choices. A response process is defined here as the cognitive processing of the item based on a specific latent person characteristic, which then affects the selection of a response category. One example of trait-unrelated response processes are response styles, which are individual preferences for specific response categories of rating scales irrespective of item content (Paulhus, 1991; Van Vaerenbergh & Thomas, 2013). Some respondents may, for instance, prefer extreme

---

<sup>1</sup>According to McReynolds and Ludwig (1987) and Ramul (1963), the first documentation of rating assessments stems from the year 1692. The German philosopher and jurist Thomasius (1692a, 1692b) assumed that every individual can be described mainly by four dimensions (rational love, sensuousness, ambition, and acquisitiveness), and he measured these on a scale with 12 categories ranging from five to 60 in steps of five. Notably, Thomasius even suggested comparing the scores of several raters (e.g., the person itself and other trained persons) in order to test the reliability of the ratings.

<sup>2</sup>This assumption is also evident in the above verses generated by ChatGPT, which resulted from a prompt asking about the purpose of rating scales. Though ChatGPT should not be considered a trustworthy source of information, I agree that rating responses can shed light on an unobservable truth.

to non-extreme categories (extreme response style; ERS), select the middle categories of a scale (midscale response style), or tend to agree with the items (acquiescence response style). Another example is socially desirable responding, which is the tendency to present oneself in an overly positive light (Paulhus, 2002). As such response processes lead to interindividual differences in the usages of rating scales that are unrelated to differences in the trait of interest, they can distort the measurement and interpretation of the results.

In addition to the variation of response processes in terms of the latent person characteristics on which they are based, other sources of heterogeneity can even further complicate the measurement of substantive traits through rating scales: First, the importance of the involved response processes for the response selection must not necessarily be homogeneous but may vary across measurement units. For instance, the items of a questionnaire may differ in how susceptible they are to influences from trait-unrelated processes such as response style-based responding. Second, heterogeneity may also exist with regard to the effects the involved processes have on the selection of response categories. Such effects are reflected in differences in the item response functions (IRFs), which represent the functional relation of the latent continuum and the response category selection. Some response processes may follow a dominance principle, meaning that higher levels of the latent person variable are associated with the selection of higher categories. Other processes may rather follow an ideal point principle, in which the relationship between the person variable and response probability is non-monotonic and depends on the proximity of person and item characteristics.

The sources of variance in the item response process described above are examples of what is referred to as heterogeneity of response processes in this thesis. More specifically, response process heterogeneity is defined as qualitative differences in the nature or composition of response processes. From the perspective of item response theory (IRT), such qualitative differences are represented by variations of structural item parameters within a given model, such as different weightings of response process dimensions (e.g., the trait or response styles) or in different parameterizations of response probabilities (e.g., through dominance or ideal point modeling). Thereby, this heterogeneity of response processes must be distinguished from the heterogeneity of respondents in terms of interindividual differences in the latent characteristics, such as traits and response styles. Variance in a person variable does not reflect qualitative differences in the item response process. Instead, it is inherent to IRT models per se and a prerequisite for item response modeling. Heterogeneity of response processes, in contrast, threatens the validity of item response models and conclusions drawn from the data if not considered in the analysis.

Therefore, modeling heterogeneity of response processes is a crucial goal in the field of psychometrics and serves two purposes: (1) Models that are based on the assumption of homogeneity may provide systematically biased estimates of the trait of interest whenever

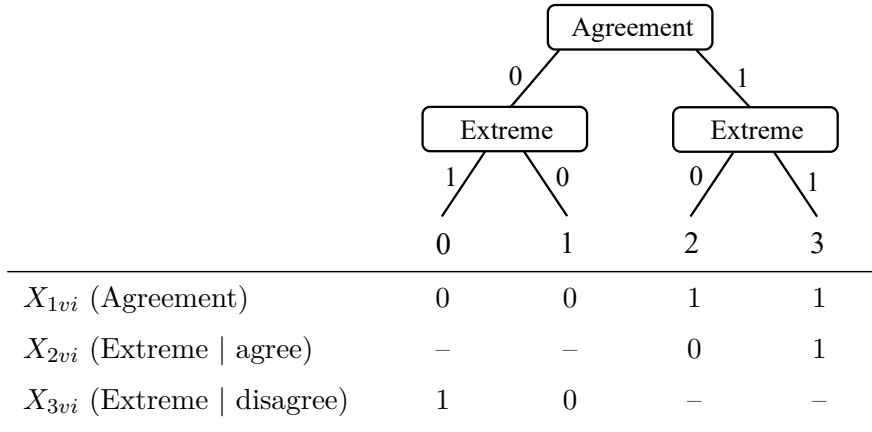
heterogeneity of response processes is present. As a result, incorporating heterogeneity facilitates accurate trait measurements. (2) Theory-driven models of heterogeneous response processes allow for an in-depth investigation of the cognitive processes that underlie item responding, and thereby help to better understand how respondents arrive at their judgments and decisions. Importantly, these two goals are closely related: A good understanding of the mechanisms behind response selections can inform the construction of analysis models and hence improve trait measurements. Good measurement models that explain the observed item response well, in turn, can give an indication of the underlying response processes. In my dissertation, I address different types of response process heterogeneity with a focus on these two goals.

The modeling framework in which the research presented here is primarily embedded in are item response tree (IRTree) models (Böckenholt, 2012; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012). The IRTree model class is based on IRT and targets multidimensional item responding. The underlying assumption of such models is that the response selection comprises several qualitatively distinct judgment steps, which are processed by the respondents according to different latent person characteristics. Therefore, IRTree models are well-suited for examining the influences of trait-based and trait-unrelated response processes and for investigating the heterogeneity of such processes.

Formally, IRTree models decompose the ordinal rating responses into pseudo-items, which represent the sub-decisions assumed to be made by the respondents during the response selection. Figure 1 illustrates an exemplary tree structure and the definition of pseudo-items for responses on a four-point rating scale, with a sub-decision of agreement and another for extreme responding conditional on the agreement judgment. The IRTree pseudo-items are parameterized by separate IRT models so that modeling choices concerning the specified response processes can be freely made for each pseudo-item and independently of the other pseudo-items. Thus, a broad spectrum of hypotheses regarding the heterogeneity of response processes – both within and between the theoretically defined processing steps – can be formulated and tested. This feature of the IRTree approach provides an advantage over other model classes tailored to multidimensional item responding, which likewise allow to control trait measurement for influences of trait-unrelated response processes: For instance, frequently applied models are multidimensional extensions of ordinal IRT models such as the multidimensional partial credit model or the multidimensional nominal response model (e.g., Bolt & Johnson, 2009; Falk & Cai, 2016; Henninger & Meiser, 2020; Wetzel & Carstensen, 2017). However, these models assume that the response selection for each item contains only one type of selection process along the latent continuum so that more fine-grained hypotheses about sub-processes cannot be specified. Further, as the definition of response processes in IRTree models relies strongly on theoretical considerations, this model class facilitates simultaneously targeting both of

**Figure 1**

*Tree Diagram and Definition of Pseudo-Items for Responses to Four-Point Rating Items*



*Note.* Pseudo-items missing by design are marked with '–'. Adapted from "Separation of traits and extreme response style in IRTree models: The role of mimicry effects for the meaningful interpretation of estimates" by V. Merhof, C. M. Böhm, and T. Meiser, 2023, *Educational and Psychological Measurement*, advance online publication, page 6, <https://doi.org/10.1177/00131644231213319>, CC BY 4.0.

the aforementioned goals of psychometrics – aiming not only to accurately measure latent constructs but also to gain insight into the cognitive processes that drive the respondents' judgments.

In this thesis, I evaluate the IRTree framework and propose further developments with the aim of improving the measurement and understanding of response process heterogeneity. The first article addresses heterogeneity with regard to the person characteristics on which the response processes are based. The focus lies on investigating the separability of the influences of multiple latent person variables, which is the prerequisite for analyzing further types of heterogeneity. I demonstrate that the separation of trait-based and trait-unrelated responding in IRTree models is at risk of being compromised under certain circumstances, and I illustrate how to detect and counteract this potential lack of validity. The second article considers heterogeneity in terms of systematically changing influences of response processes on responding to the items over the course of a questionnaire. I develop dynamic IRTree models that account for such variations of process involvements and quantify trajectories of the respondents' strategies over time. The third article is concerned with the heterogeneity of how the response processes influence the selection of rating categories, that is, the type of their IRF. I propose a multidimensional IRT (MIRT) model, with which the combined effects of both dominance and ideal point processes can be accommodated. Further, I demonstrate how the scope of IRTree modeling is broadened

by using the new model for the parameterization of pseudo-items.

Throughout this thesis and the included articles, several examples from the field of response style modeling are presented. These illustrations were chosen to account for the fact that a large body of research on analyzing trait-unrelated response processes focuses on response styles. This is not surprising since response styles jeopardize the reliability and validity of both individual assessments and group comparisons, and thereby may ultimately compromise the utility of all kinds of measurements through rating scales (Baumgartner & Steenkamp, 2001; Cheung & Rensvold, 2000; Morren et al., 2012). However, it is important to note that other types of response processes can distort trait measurement as well and are likewise the target of model-based approaches in the literature. Socially desirable responding, for instance, is a great concern mainly of high-stakes assessments and has been investigated in numerous studies (e.g., LaHuis & Copeland, 2009; Leng et al., 2020; Sun et al., 2022; Ziegler & Buehner, 2009). In addition, there exist various other models that address item responding in specific measurement contexts and consider, for example, preference for fast versus accurate responding (in the sense of the speed-accuracy trade-off) or the propensity to omit items (e.g., Maris & van der Maas, 2012; Ulitzsch et al., 2020; van Rijn & Ali, 2018). The modeling approaches presented here are therefore intended to illustrate how heterogeneous response processes can be studied in the context of response style modeling, and may provide starting points for further generalizations.

The next chapter gives an overview of IRTree parameterizations for different response processes as well as their heterogeneity, and addresses the issue of separating multiple processes relating to different person characteristics. Then, two forms of response process heterogeneity are discussed in more detail, and IRTree model extensions are proposed to account for such. An empirical application demonstrates how the presented approaches can be used to improve the measurement of response process heterogeneity in practice and to gain new insights into the cognitive processing of items. Lastly, the implications of the presented research are discussed, limitations are identified, and directions for future research are derived.





## 2 IRTree Modeling

IRTree models assume that responding to rating items gives rise to multiple cognitive processing steps. This theoretical concept is implemented in the models by the decomposition of the ordinal rating responses  $Y_{vi} \in \{0, \dots, K\}$  of person  $v = 1, \dots, N$  to item  $i = 1, \dots, I$  into a sequence of pseudo-item responses  $X_{hvi}$ . The probability of an ordinal response is then obtained by the product of the probabilities of responses to the respective pseudo-items. Importantly, the defined sequence of such pseudo-items refers to a logical conditionality of judgment steps and does not necessarily imply a temporal sequence.

For instance, the exemplary IRTree model depicted in Figure 1 consists of two sub-decisions, one reflecting agreement and another one reflecting extreme responding conditional on the agreement judgment. Therefore, the probability of the ordinal responses  $Y_{vi} \in \{0, \dots, 3\}$  is derived from the pseudo-item responses  $X_{hvi} \in \{0, 1\}$  by

$$p(Y_{vi} = y_{vi}) = p(X_{1vi} = x_{1vi}) \times p(X_{2vi} = x_{2vi})^{x_{1vi}} \times p(X_{3vi} = x_{3vi})^{(1-x_{1vi})}, \quad (2.1)$$

where  $X_{1vi}$  denotes the agreement pseudo-item response ( $h = 1$ ) and  $X_{2vi}$  and  $X_{3vi}$  denote the responses to the two extreme pseudo-items conditional on agreement and disagreement, respectively ( $h = 2$  and  $h = 3$ ).

In applications to response style modeling, a commonly made assumption is that agreement judgments relate to a trait-based response process, whereas all other judgments reflect responding based on response styles. For the IRTree structure described above, this rationale can be translated into applying a unidimensional IRT model of the substantive trait  $\theta$  to the agreement pseudo-item and unidimensional models of the ERS factor  $\eta$  to the two extreme pseudo-items. Under a Rasch parameterization, the probabilities of the pseudo-item responses can then be obtained by:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})}, \quad (2.2)$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\eta_v - \beta_{i2}))}{1 + \exp(\eta_v - \beta_{i2})}, \quad (2.3)$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\eta_v - \beta_{i3}))}{1 + \exp(\eta_v - \beta_{i3})}, \quad (2.4)$$

where  $\beta_{ih}$  denotes the difficulty of pseudo-item  $h$  of item  $i$ .

This parameterization implies a rather simplistic way of item processing and can be considered one of the most basic forms of an IRTree model. Such a model assumes a high degree of response process homogeneity (e.g., the influences of trait and ERS are assumed to be constant across items) and allows to examine heterogeneous response processes to a very limited extent (e.g., whether the pseudo-item difficulty differs for extreme responding conditional on agreement and disagreement). The IRTree model class can, however, be implemented with more complex parameterizations and can incorporate manifold types of heterogeneity. The following section elaborates on how different forms of heterogeneity can be included and gives a systematic overview of modeling choices in the IRTree framework.

## 2.1 Parameterizations of Response Processes and their Heterogeneity

IRTree models specify the involvement and heterogeneity of response processes on the level of a priori defined processing stages. Thereby, they can be regarded as a modular system, in which different modeling components can be freely combined: The first component is the psychological theory about the sub-decisions involved in the selection of ordinal categories and the conditionality of the cognitive judgment steps. The second component is the assignment of the response processes to the individual pseudo-items, which represent such sub-decisions. The third component is the parameterization of the pseudo-items as a function of the assigned response processes. Each of these components allows to integrate various assumptions on the response processes, and thus, to formalize and study different forms of heterogeneity.

The first component, the partitioning of the ordinal responses into sub-decisions, defines the hypothesized level of granularity of the judgments. Typically, IRTree models consist of binary sub-decisions only, though they can likewise include judgments with three or more options (see Meiser et al., 2019). However, the fewer qualitatively distinct sub-decisions are defined, the lower the assumed complexity of the item response process, and the lower the capability of the model to accommodate heterogeneous response processes. Furthermore, the partitioning of the ordinal responses also determines whether the tree structure is symmetrical or asymmetrical. In symmetrical models, the same sequence of sub-decisions is assumed to underlie the selection of corresponding categories on both sides of the rating scale (i.e., categories 0 and  $K$ , 1 and  $K - 1$ , and so on). In contrast, asymmetrical structures reflect the assumption that the selection of corresponding categories is driven by different sequences of sub-decisions (for an overview of different kinds of IRTree structures, see Jeon & De Boeck, 2016). Nonetheless, asymmetrical models do not necessarily imply a higher complexity of the response selection: Sequential models, for instance, presume that

respondents decide in successive steps whether to select either a specific category or one of the higher categories, so that the response options are evaluated in ascending order (Tutz, 1990; Verhelst et al., 1997). Such models are asymmetrical in structure but could still be homogeneous in the sense that the entire selection process is a unidimensional trait-based one. The degree of heterogeneity incorporated in an IRTree model is consequently affected but not inherently predetermined by the definition of the tree structure.

The second IRTree component, the assignment of response processes to the pseudo-items, specifies which person characteristics are assumed to be involved in each processing stage. In simple structure IRTree models, all pseudo-items are defined to be dependent on one response process each, so the multidimensionality arises only between pseudo-items. More complex, multidimensional definitions of pseudo-items reflect the hypothesis that a sub-decision depends on multiple response processes simultaneously (Böckenholt, 2019; Jeon & De Boeck, 2016; Meiser et al., 2019).<sup>3</sup> The choice of the pseudo-item dimensionality and the flexible assignment of response processes, therefore, facilitates incorporating heterogeneity with regard to the type of underlying person characteristic both within and between pseudo-items.

The third component, the IRT models used to parameterize the pseudo-items, defines how each of the response processes influences the response selection. Depending on how many options are assigned to the respective pseudo-item, various models for binary or ordinal data are available, and depending on how many response processes are assigned, unidimensional or multidimensional models can be selected accordingly. Heterogeneity of response processes can thereby be integrated in many different ways: For example, models with item-specific discrimination parameters allow for varying impacts of the respective person characteristics throughout a questionnaire. Commonly used models are the 2PL model (Birnbaum, 1968), the generalized partial credit model (Muraki, 1992), or the graded response model (Samejima, 1969). In contrast, models with fixed item discrimination reflect the assumption of a constant impact across items, such as the Rasch model (Rasch, 1960), the one-parameter probit model (Birnbaum, 1968), or the partial credit model (Masters, 1982). More recently, IRTree pseudo-items have been parameterized by ideal point models, which differ from the aforementioned models belonging to the group of dominance models in how the levels of latent person characteristics are mapped to the item scores. Examples of ideal point models are the (generalized) graded unfolding model (Roberts et al., 2000; Roberts & Laughlin, 1996) or the (generalized) hyperbolic cosine model (Andrich, 1996; Andrich & Luo, 1993). The choice of the IRT model used

---

<sup>3</sup>IRTree models are sometimes described as multi-process models, where a *process* is then conceptualized as a judgment step during the response selection (e.g., the process of deciding on whether to agree or disagree). This definition is to be distinguished from the definition of a *response process* used here, which refers to making a judgment on the basis of a person characteristic. Though the two conceptualizations overlap in simple structure IRTree models, they do not necessarily do so in models with multidimensional pseudo-items.

to parameterize the IRTree pseudo-items (e.g., dominance versus ideal point IRF; with versus without discrimination parameters) thus determines the degree of heterogeneity within each single processing step.

However, despite the great flexibility of the IRTree framework in all three components, it is important to note that the higher the heterogeneity built into the model, the more complex the estimation becomes, and the higher the risk of identifiability problems. Researchers should, therefore, avoid applying highly parameterized models solely for the sake of improving model fit and instead specify their models based on theoretical considerations. Thereby, it may be reasonable (or even necessary) to impose parameter constraints that reduce the model's complexity while still reflecting the hypotheses on the item processing as closely as possible. For instance, a typical assumption in the literature is that two corresponding pseudo-items, which refer to the same theoretical sub-decision (e.g., extreme responding conditional on agreement and disagreement), are parameterized by the same set of person parameters. Often, an even stronger invariance assumption is made (called directional invariance, see Jeon & De Boeck, 2019), in which the thresholds and/or discrimination parameters of such pseudo-items are set equal (e.g., Böckenholt, 2017; Jin et al., 2022; Kim & Bolt, 2021; Plieninger, 2020). These and other modeling choices should be thoroughly considered and adapted in a way that they balance the practical aspects of the model estimation on the one hand, and the capability to account for the potential heterogeneity of response processes on the other.

An IRTree parameterization that is frequently referred to in the articles and in this thesis is an extension of the model defined above (Equation 2.2 to 2.4) by multidimensional extreme pseudo-items. The judgments of extreme responding are assumed to not only be driven by the ERS  $\eta$  but additionally by the substantive trait  $\theta$ . Given that high trait levels can be expected to increase the probability of choosing high categories, trait-based extreme responding is assigned a positive influence conditional on agreement and a negative one conditional on disagreement. One possible implementation of these assumptions yields the following parameterization:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})}, \quad (2.5)$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\eta_v + \alpha\theta_v - \beta_{i2}))}{1 + \exp(\eta_v + \alpha\theta_v - \beta_{i2})}, \quad (2.6)$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\eta_v - \alpha\theta_v - \beta_{i3}))}{1 + \exp(\eta_v - \alpha\theta_v - \beta_{i3})}, \quad (2.7)$$

with  $\alpha \geq 0$ . In the articles of this thesis, different modified versions of this parameterization were used, among others, by including item-specific influences of trait and ERS or

by specifying ideal point response processes.

## 2.2 Separation of Response Processes

Merhof, V., Böhm, C. M., & Meiser, T. (2023). Separation of traits and extreme response style in IRTree models: The role of mimicry effects for the meaningful interpretation of estimates. *Educational and Psychological Measurement*. Advance online publication. <https://doi.org/10.1177/00131644231213319>

In order to investigate and account for response process heterogeneity in IRTree models, the influences of such processes on the respondents' category selection must be separated from each other. Importantly, both the statistical and the meaningful separation of processes have to be ensured. The first one is given if the model and the person variables are identified. The second one additionally requires that the substantive meanings of the response processes are distinct and do not overlap, which is the case if the specified effects of the person parameters on the selection of ordinal categories cannot be linearly transformed into each other. In the IRTree framework, however, the meaningful separation of response processes based on the trait and the ERS may be compromised by a so-called mimicry effect: The response style factor then mimics part of the substantive trait and captures variance in item responding induced by a trait-based response process. The simulation studies presented in the first article of this thesis demonstrated that mimicry effects result in inflated estimates of the ERS variance and a biased estimation of the covariance between response style and trait. Accordingly, false conclusions may be drawn regarding the impact of the ERS on the respondents' judgments as well as the relationship between individual category preferences and the levels of the measured construct.

Notably, the simulation studies also revealed that mimicry effects can only occur if two conditions are given: First, the ordinal item responses have to be asymmetrically distributed across the categories of the rating scale. Such an asymmetry can be caused, for instance, by a distribution of the respondents' trait levels which is skewed or shifted in relation to the distribution of items.<sup>4</sup> Thereby, the higher the asymmetry of ordinal responses, the more trait-induced variance can be captured by the ERS factor, and the stronger the mimicry effect becomes. Second, mimicry effects only arise if an IRTree model is misspecified in a way that the influence of the response style is overstated while the influence of the trait is understated. This is the case, for example, if judgments of extreme responding are modeled to be dependent on the ERS, when respondents actually derived their answers (additionally or exclusively) by a trait-based process. As a consequence,

---

<sup>4</sup>Note that the less balanced positively and negatively keyed items are in a questionnaire, the stronger the asymmetry of responses evoked by shifted or skewed trait distributions.

simple structure IRTree models with unidimensional pseudo-items were shown to carry a high risk of mimicry effects, as the assumption that only one response process each affects the respondents' sub-decisions is likely a simplification of the true data-generating process in real-world applications. In one of the empirical examples presented in the article, for instance, an indication for a substantial mimicry effect in a simple structure model was found, as the results suggested that the ERS variance as well as the correlation between trait and ERS were considerably overestimated (the model's estimates for the variance and correlation were 5.74 and 0.41, respectively; another analysis indicated that the variance and correlation actually were 4.02 and 0.19, respectively; see Table 7 on page 21 of the first article of this thesis). Model specifications with multidimensional pseudo-items, in contrast, cover a wider range of plausible ways of item response processing. Such models were found to be less susceptible to mimicry effects and better suited to separate trait and response style factors irrespective of the distribution of ordinal responses. Furthermore, it was revealed that even if pseudo-items were overparameterized and contained response processes that were actually not relevant for the response selection, the models reliably reflected the absence of such a process.

In light of these findings, traditional IRTree models with unidimensional pseudo-items should be applied with caution, and researchers should be aware of the possibility that the estimated parameters may not have the substantial meaning assigned to them when specifying the model. However, it was also shown that the estimation of the substantive trait levels was barely impaired by mimicry effects, as the rank order of respondents could be very well recovered even by misspecified IRTree models. This demonstrates that the potential misattribution of variance components is a more or less severe problem for IRTree analyses depending on which of the two aforementioned purposes of psychometric modeling is pursued: While the aim of accurate trait measurements is hardly affected by mimicry effects, the investigation of underlying cognitive processes may be severely compromised. Accordingly, simple structure IRTree models may be sufficient for certain practical applications, for example, if potential response styles are treated as nuisance factors that are included solely for the sake of unbiasing trait estimates. Nevertheless, it seems worth considering IRTree models with more complex parameterizations whenever the response styles themselves or the cognitive response strategies are of interest. In addition, it is not clear to what extent such misspecified IRTree models may distort the trait estimation under other circumstances not covered by the simulation studies. It thus seems advisable to rather take additional processes into account, even though they may potentially be irrelevant, instead of running the risk of neglecting one.

This suggestion applies all the more since the impaired separability of response processes may not only distort our understanding of the respondents' behavior in empirical applications – it also poses a threat to the validity of simulation studies aiming to systematically

investigate specific properties of IRTree models. A common procedure for generating item response data in simulation studies is to sample person-specific trait levels and item difficulty parameters from symmetrical distributions that have the same mean structure (such as normal or uniform distributions centered around zero, e.g., Kim & Bolt, 2021; Leventhal, 2019). The data generated in this way precludes the occurrence of mimicry effects, as the resulting ordinal response distributions are largely symmetrical (with small variations due to random sampling). It is, therefore, unclear to what extent the results of the simulation studies also hold in the presence of asymmetrical distributions, which potentially give rise to mimicry effects, or whether such would change the conclusions drawn from the studies.

Further research is also needed to examine the generalizability of the findings on mimicry effects obtained from the simulation studies beyond the specific conditions covered in the article. For instance, it should be clarified how the mimicry effect addressed there, in which the ERS factor mimics the trait, can be transferred to other response processes (e.g., based on midscale or acquiescence response style). Although logical considerations suggest that mimicry effects should have similar consequences and remedies for other response styles, it is unclear how they would manifest if multiple response styles simultaneously affected the category selection or if other trait-unrelated processes such as socially desirable responding were present. In addition, though multidimensional pseudo-items have been shown to counteract mimicry effects under certain kinds of data-generating procedures (e.g., unidimensional data without response style influence), they may still be misspecified and lead to mimicry effects under other conditions – for instance, if further response processes are involved in the data generation but are not considered in the analysis.

Despite this critical view on the validity of some IRTree model specifications, however, it should be emphasized that the question of meaningful model parameters is not unique to IRTree modeling: Mimicry effects are likely to also occur in other IRT model classes designed to control for response style effects, such as the multidimensional nominal response model (e.g., Bolt et al., 2014; Falk & Cai, 2016; T. R. Johnson & Bolt, 2010). Moreover, the potential lack of separability of factors addressed here has some parallels to estimation issues and interpretability problems in confirmatory factor analysis. In G-factor models (such as the bifactor model), for example, it is recognized that the interpretation of the variance components assigned to the general factor and to the specific factors can be difficult (Eid et al., 2017). Likewise, multitrait-multimethod models are prone to interpretation issues, in particular when the factors are correlated (Eid, 2000; Eid et al., 2003). This shows that the challenge of separating response processes is neither specific to IRTree nor IRT analyses, and that the substantive meaning of model parameters is a general concern of psychometrics, which should be paid more attention to in future research.





### 3 Modeling Heterogeneity of Response Processes

The previous section introduced the arguably most fundamental and most frequently investigated form of response process heterogeneity, which is the heterogeneity in terms of the person characteristics the processes depend on. There exists a wide range of modeling approaches within and outside the IRTree class that address such heterogeneity by not only considering the substantive trait, but additionally incorporating trait-unrelated influences such as response styles (e.g., Debeer et al., 2017; Falk & Cai, 2016; Khorramdel et al., 2019; Leng et al., 2020; Plieninger & Heck, 2018; Thissen-Roe & Thissen, 2013; Wetzel et al., 2013). However, various other manifestations of heterogeneous response processes may arise, two types of which are elaborated in the following: In the first part, dynamic within-person heterogeneity is discussed, in which the importance of the involved response processes varies over the course of the questionnaire. In the second part, heterogeneity with respect to the IRFs of multiple processes is taken into account.

Importantly, the occurrence of these forms of heterogeneity can be expected to depend on the measurement instrument and situation. Relevant factors may be, among others, the measured construct, the wording of the items, the arrangement of items in the questionnaire, or whether the assessment is high or low stakes. Consequently, the resulting response process heterogeneity must be distinguished from heterogeneity that is not due to the context: For example, respondents may generally differ in what kind of response processes they use to what extent when generating item responses, causing between-person heterogeneity. Likewise, the items of a questionnaire may differ in how respondents perceive and process them, resulting in between-item heterogeneity. Investigating such is beyond the scope of this dissertation, though the modeling approaches introduced here could potentially be extended to incorporate these types of heterogeneity as well, as will be discussed in later sections.

### 3.1 Dynamic Within-Person Heterogeneity of Response Process Involvement

Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, 88(4), 1354-1380. <https://doi.org/10.1007/s11336-023-09901-0>

A within-person heterogeneity of the response process involvement is present when the influences of such processes on the respondents' category selection vary across items. If this variability exhibits a systematic change throughout a questionnaire, it is referred to as dynamic heterogeneity. Such a dynamic response strategy may occur, for example, if the respondents' motivation decreases over time, so that heuristic response processes like response style-based responding may gain influence.

In order to account for systematic trajectories of the involvement of trait-based and response style-based processes, dynamic IRTree models were developed in the second article of this thesis. Such models define item position-dependent loadings of the corresponding person variables.<sup>5</sup> A continuous version of the dynamic models constrains the loadings by a continuous linear or curvilinear trajectory and thereby captures the underlying trend of the response strategy in a parsimonious way. A second, more flexible version of the dynamic IRTree approach allows for additional random fluctuations of the loadings and can reveal item-specific effects beyond the item position (see Figure 2 for an illustration of the two model versions).

In both the continuous and the flexible model, dynamic changes can be incorporated in unidimensional and multidimensional pseudo-items, whereby the trajectories are specified for each of the processes separately. Simulation studies showed that dynamic IRTree models accurately quantify systematically changing influences of response processes across the items of a questionnaire. Further, they also detect the absence of dynamics, that is, response processes that have a constant influence across items. Thus, dynamic models are well suited to analyze the respondents' behavior and are a valuable contribution with regard to the goal of deepening our understanding of the cognitive processes underlying item responding.

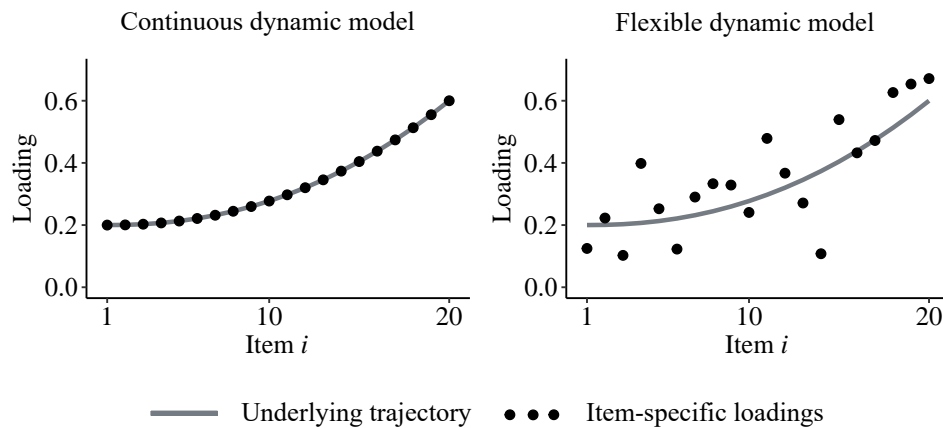
The empirical example in the article indeed demonstrated that new insights into the item response process can be gained by dynamic modeling: For the analyzed data set, the sub-decisions differed in the degree to which systematic changes occurred, and trait-based and response style-based responding differed in the amount of unsystematic variation of

---

<sup>5</sup>The term *loading* is used here to describe the weight of a person variable and therefore differs from the definition of a discrimination parameter, which weights the difference between the person variable and the item difficulty.

**Figure 2**

*Exemplary Dynamic Trajectories and Item-Specific Loadings Under the Continuous Dynamic Model and the Flexible Dynamic Model*



*Note.* Adapted from "Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes" by V. Merhof and T. Meiser, 2023, *Psychometrika*, 88(4), page 1370, <https://doi.org/10.1007/s11336-023-09901-0>, CC BY 4.0.

the loadings. This suggests that the hypothesized judgment steps evoked qualitatively different ways of cognitive processing. Building on these preliminary findings, future research could investigate whether such heterogeneity of response processes is a general phenomenon and can be found in other empirical data sets as well.

In addition, models of heterogeneous response processes provide a means to put theoretical expectations of homogeneity to the test, by analyzing the effects of certain parameter constraints. For example, constraints on dynamic trait trajectories could be imposed in order to explore whether the involvement of trait-based processes followed similar patterns across sub-decisions (e.g., for agreement and extreme responding). Another reasonable constraint of dynamic models could be that trajectories of trait-based and response style-based processes are defined as mutually dependent with opposite directions. This would allow examining whether a decreasing influence of the substantive trait was accompanied by a correspondingly greater impact of response style-based processes, as one might expect if respondents became fatigued or impatient over time.

It is important to keep in mind, however, that when investigating dynamic response strategies in such an exploratory way, only monotonous changes can be detected. As a result, more complex assumptions, such as U-shaped trajectories, cannot be tested by the dynamic IRTree models as proposed in the article. Though the monotonous function used to define the process loadings could, in principle, be replaced by any other function, the estimation of non-monotonous trajectories may be challenging. Another limitation with

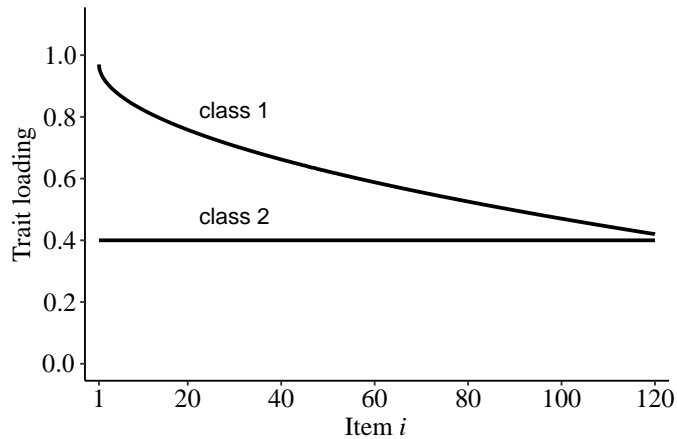
regard to the modeling of trajectories is that the dynamic changes are assumed to apply equally to all respondents. It is, however, very likely that the causes of dynamic strategies, such as changes in test-taking effort, additionally vary across respondents (e.g., if respondents differ in their interest in the measured construct or if they experience different levels of time pressure while participating in the survey). It thus seems reasonable to further extend the dynamic IRTree models in a way that between-person heterogeneity can be accounted for. Therefore, the dynamic IRTree approach was combined with mixture modeling for an analysis presented at the International Meeting of the Psychometric Society 2022 (Merhof & Meiser, 2022). The application to an empirical data set demonstrated that, indeed, respondents may show different patterns of dynamic response behaviors: In the exemplary data set, a decrease in trait-based responding accompanied by an increase in response style-based responding was found for one latent class of respondents, whereas another class rather used a constant response strategy (see Figure 3 for the trait loading trajectories estimated by the continuous dynamic IRTree model extended by a person-mixture). Knowledge about such groups of respondents with different response strategies can inform test construction and thereby potentially improve the data quality.

Despite the advantages and promising further developments of the dynamic approach for investigating the respondents' behavior, however, the findings from the simulation studies reported in the article revealed that dynamic modeling only slightly benefits the trait estimation: The recovery of latent trait levels was hardly affected if the response process involvement systematically varied in the data but was ignored by an analysis model that assumed constant influences of all processes. As was already discussed above in the context of the mimicry effect, misspecified IRTree models seem to be of less concern if the sole purpose of an analysis is the trait measurement without intending to draw conclusions on the cognitive processes underlying the responses.

Still, the simulation studies also suggested that not all types of misspecifications lead to equally accurate parameter estimates (of person and of item parameters) under all circumstances. More specifically, it was relevant only to a limited extent whether the influences of the response processes were correctly defined for all items, that is, whether dynamic trajectories or constant loadings were modeled. Instead, it appeared to be of higher importance that all processes involved in the category selection were actually included in the parameterization of the corresponding pseudo-items. Accordingly, simple structure IRTree models in particular were found to be at risk of providing inaccurate estimates whenever the true data-generating process differed from the model-implied one. IRTree models with multidimensional pseudo-items, in contrast, produced accurate estimates even if they were overparameterized and incorporated processes that were not actually involved in the judgments. These findings corroborate the previous recommendation that in case of uncertainty about the complexity of the item response process,

**Figure 3**

*Person-Mixture of Dynamic Trait Loading Trajectories in Empirical Data*



*Note.* The two trajectories represent the trait loadings of the extreme pseudo-items estimated by the continuous dynamic IRTree model extended by a person-mixture. 39% of the respondents are estimated to belong to class 1 and 61% to class 2. The analyzed data set stems from J. A. Johnson (2014) and consists of item responses to the IPIP-NEO-120 personality inventory with 120 items. More information on the data set and the data preparation can be found in the second article of this thesis, in which the same data set was used as an empirical application example.

researchers should rather include additional processes into IRTree pseudo-items instead of being at risk of missing one.

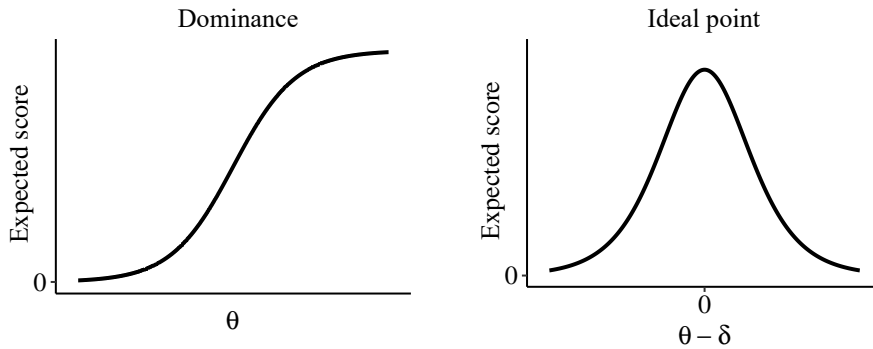
### 3.2 Heterogeneity of Item Response Functions

Merhof, V., & Meiser, T. (2023). *Co-occurring dominance and ideal point response processes: A general IRTree framework for multidimensional item responding*. Revision invited by Behavior Research Methods.

The response processes involved in item responding can be further heterogeneous in the sense that they adhere to different item response functions (IRFs). An IRF defines how the values of the respective latent person variable are mapped to the expected score of an item. IRTree pseudo-items are traditionally and most frequently parameterized by dominance models, such as the Rasch or 2PL model. Dominance IRFs are monotonous and reflect the assumption that higher levels of a person characteristic correspond to higher expected scores (see Figure 4, left panel). More recently, IRTree pseudo-items have been parameterized by ideal point models (also called unfolding models), which assume that

**Figure 4**

*Item Response Functions Under the Dominance and Ideal Point Assumption*



*Note.* The parameter  $\theta$  denotes the substantive trait, and  $\delta$  denotes the item location under the ideal point model. Adapted from "Co-occurring dominance and ideal point response processes: A general IRTree framework for multidimensional item responding" by V. Merhof and T. Meiser, 2023, page 6.

the relationship between the latent characteristic and the expected score is unimodal and non-monotonic (see Figure 4, right panel). The expected score is highest if the person's level of the latent characteristic (i.e., their ideal point) matches the location of an item and decreases with greater distances. The more a person's ideal point deviates from the item location in an upward or downward manner, the more likely he or she disagrees with the item, referred to as disagreement from above and below, respectively. The ideal point rationale was found to often better describe self-reported responses to attitudinal items and other non-cognitive constructs (for overviews, see Drasgow et al., 2010; Tay & Ng, 2018). Given that IRTree models are usually applied to this kind of data, it seems reasonable to make use of ideal point parameterizations more frequently.

By using existing IRT models, however, ideal point response processes can only be implemented for unidimensional pseudo-items, while multidimensional pseudo-items are limited to dominance models. Therefore, in the third article of this thesis, a general MIRT model of co-occurring processes was developed, which facilitates accounting for the combined involvement of all kinds of processes within a single pseudo-item (e.g., if an ideal point trait and a dominance response style are both assumed to affect a sub-decision). IRTree pseudo-items can be parameterized by the new MIRT model in a consistent way, and multiple dominance and ideal point processes can be modeled to affect the response selection both sequentially across pseudo-items and as co-occurring processes within pseudo-items.

Two application examples confirmed that dominance and ideal point processes may be simultaneously involved in the respondents' judgments in real data. One example demon-

strated the effectiveness of the new IRTree parameterization for controlling trait measurements through ideal point items (i.e., trait-based responding followed the ideal point assumption) for ERS. As response style-based responding by definition is a dominance process, modeling trait and ERS influences by traditional IRTree parameterizations was restricted to unidimensional pseudo-items. In contrast, the proposed MIRT model allowed the inclusion of simultaneous effects of both processes in multidimensional pseudo-items – which, in fact, turned out to be advantageous for the given data set. A second empirical example showed that co-occurring dominance and ideal point processes can also be beneficial for analyzing items that follow the dominance assumption, namely when modeling sub-decisions of midscale versus non-midscale responding: Since respondents with rather extreme trait levels in relation to the item location can be assumed to have a clear-cut opinion on the item content, and only respondents with moderate trait levels relative to the item location are expected to select middle categories as an expression of a neutral opinion, unimodal ideal point IRFs are well suited for describing trait-based midscale responding. If in addition to such an ideal point process, also the respondents' midscale response style (i.e., a dominance response process) is to be considered in midscale sub-decisions, a dominance and an ideal point process co-occur. The empirical example in the article did not only provide evidence that such a modeling approach fits the data better than alternative ones: By analyzing item-level response time data in addition to the item responses, support for the construct validation for parameters of the new IRTree parameterization was found.

A simulation study further revealed that IRTree models with the proposed parameterization can recover person and item parameters well and successfully capture the co-occurrence of an ideal point trait and a dominance response style. In contrast, if one of the co-occurring response processes was ignored and an IRTree model with unidimensional pseudo-items was falsely applied, larger estimation errors resulted. Moreover, it was shown that model fit comparisons were well suited to determine whether the data-generating process incorporated both a dominance and an ideal point process or whether one of the processes was actually not involved. Thus, using the MIRT parameterization of co-occurring processes was found to serve both psychometric purposes; the trait measurement as well as the investigation of IRFs of response processes, which in turn provides information about the cognitive processes involved in item responding. Thereby, the conducted work contributes to an ongoing discussion in the psychometric literature on whether non-cognitive constructs are generally better described by ideal point models, and under which conditions dominance models are appropriate (e.g., Chernyshenko et al., 2007; Drasgow et al., 2010; Stark et al., 2006; Tay & Ng, 2018). With the new approach, these questions can be addressed while taking the distorting influences of response styles into account.

Beyond that, further applications outside of response style modeling are straightforward: In an empirical application conducted for a poster presentation (Merhof & Meiser, 2023), it was shown that item omissions could be well described by a MIRT parameterization in which the respondents' omission propensity and the substantive trait co-occurred.<sup>6</sup> Figure 5 illustrates the prediction of item omissions under such a MIRT model of co-occurring processes. Similar to modeling middle categories in dominance items, the trait was assumed to behave like an ideal point process despite the fact that the items of the questionnaire followed the dominance rationale: Respondents with very high or very low trait levels in relation to an item's location were expected to have a clear-cut opinion on the item content and, therefore, to have a low propensity to omit this item. Respondents with moderate trait levels relative to the item location, in contrast, were expected to be rather undecided and more likely to skip the item. Indeed, trait-based responding was found to follow the ideal point rationale in the analyzed data set, and an IRTree model of co-occurring processes (of the ideal point trait and the dominance omission propensity) captured the response behavior well.

Another conceivable further development of the new IRTree approach is the combination with mixture modeling. On the one hand, including a person-mixture seems promising, as respondents may differ in how they interpret the items and whether their trait-based response selection followed the dominance or ideal point rationale. On the other hand, there might also be a mixture of items within a questionnaire, where part of the items could evoke either dominance or ideal point responding (e.g., due to differences in item wording; also see Weekers & Meijer, 2008). However, future research would be needed to examine whether such an additional level of modeling complexity still can be realized in practice and whether sufficiently accurate parameter estimates can be obtained.

New directions for future research may also arise from linking ideal point IRFs with the modeling approaches presented in the other two articles: For example, it could be reasonable to assume that dynamically changing influences, as presented in the second article of this thesis, also apply to ideal point processes. Since the MIRT parameterization of co-occurring processes allows to estimate item-specific loadings for each of the processes, a constraint of such to continuous trajectories should be straightforward and easy to implement.

Also in relation to the first article concerning the separation of trait and extreme response style factors, more research on ideal point response processes might provide interesting insights. An unpublished small simulation study suggested that mimicry effects likewise occur if trait-based responding followed an ideal point instead of dominance IRF. Interestingly, however, the characteristics and consequences of the mimicry effects differ:

---

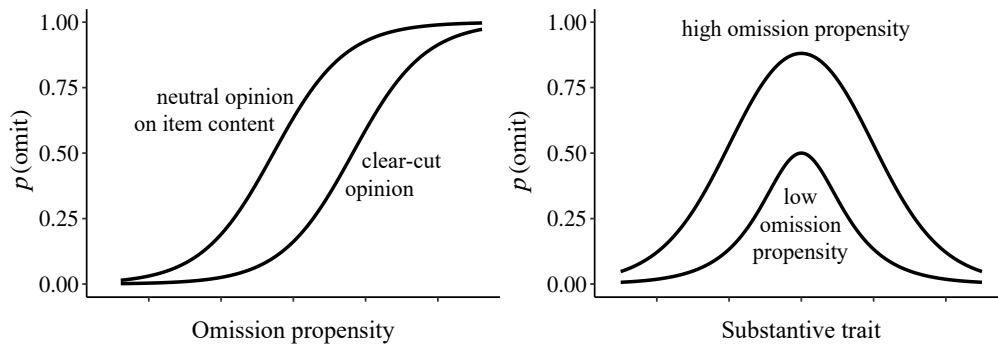
<sup>6</sup>The analyzed data was a subset of the data collected in the Trends in Mathematics and Science Study from the year 2019. Item responses of eighth-grade students to part of the student context questionnaire were used.



**Figure 5**

*Probability of Item Omissions Under a MIRT Model of Co-Occurring Dominance*

*Omission Propensity and Ideal Point Trait*



Under the dominance assumption, a trait distribution that is shifted toward high trait levels in relation to the item distribution results in an overestimation of the relationship between trait and ERS, while this relationship is underestimated for shifts toward low trait levels. For ideal point traits, it is the other way around, so that shifts toward high and low trait levels relative to the distribution of item locations lead to underestimating and overestimating the relationship, respectively. In addition, and in contrast to dominance traits, both directions of trait shifts result in an asymmetrical distribution of ordinal item responses toward lower categories of the scale (as a result of disagreement from below and from above). Thus, it is not possible to infer from the observed response distribution whether a mimicry effect would rather be reflected in an over- or underestimation of the relationship between trait and ERS, so mimicry effects are probably even more difficult to detect when applying ideal point modeling. Nonetheless, mimicry effects can be expected to be overcome by multidimensional pseudo-items also for ideal point traits, which underlines the benefits of the new parameterization of co-occurring processes.



## 4 Illustrative Application

This chapter provides an empirical application example and demonstrates how the heterogeneity of response processes can be studied by using the IRTree models presented in the articles included in this dissertation. The analyzed data set stems from the Trends in Mathematics and Science Study (TIMSS) from the year 2019. Item responses of 2,503 German fourth-grade students to a scale of the context questionnaire were used. The scale measures the students' attitude toward learning mathematics by nine items on a four-point rating scale.

Appendix A gives more information on the data set and the data preparation. In addition, the Bayesian model estimation and the model comparison by an approximation of leave-one-out cross-validation (LOO) are described there. Further, the parameterization of pseudo-items as well as prior choices for the applied IRTree models are reported. All models are based on the tree structure depicted in Figure 1.

### 4.1 Separation of Trait and Extreme Response Style

The first analysis focused on the separation of the measured substantive trait (i.e., the students' attitude) and ERS. To this end, two IRTree models were fitted to the data set: the simple structure model with unidimensional pseudo items (trait-based agreement and ERS-based extreme responding; see Equation 2.2 to 2.4) and an extended model with multidimensional extreme pseudo-items (dependent on both trait and ERS; see Equation 2.5 to 2.7). The results are summarized in Table 1.

The comparison of these models by the LOO information criterion (LOO IC) demonstrated that the respondents' judgments of extreme responding were not only influenced by an ERS-based response process but additionally by a trait-based one. This, in turn, indicates that applying the model with unidimensional pseudo-items potentially entails the risk of a mimicry effect, since the influences of the response style and the trait are likely to be overstated and understated by the model, respectively. Indeed, indications of a pronounced mimicry effect were found, as the estimated correlation between ERS and trait was of substantial size for the model with unidimensional pseudo-items but negligible for the model with multidimensional ones. Furthermore, the large difference in the estimated ERS variance between the models likewise revealed a strong effect. The results, therefore, provide evidence for impaired separability of response processes by the model

**Table 1***Mimicry Effect in TIMSS 2019 Data*

Extreme responding	LOO IC	Correlation & variances of ERS $\eta$ and trait $\theta$		
		$\widehat{\text{Cor}}(\eta, \theta)$	$\widehat{\text{Var}}(\eta)$	$\widehat{\text{Var}}(\theta)$
Unidimensional	37,806	0.38 [0.33, 0.42]	5.44 [4.96, 5.93]	8.75 [7.93, 9.63]
Multidimensional	36,510	0.08 [0.00, 0.16]	2.97 [2.65, 3.32]	9.15 [8.27, 10.05]

*Note.* 95% credible intervals in brackets.

with unidimensional pseudo-items and suggest that multidimensional pseudo-items should be used for drawing substantive conclusions from the data.

Accordingly, also the heterogeneity of response processes was evaluated based on the estimates of the model with multidimensional pseudo-items: First, the response processes involved in the respondents' judgments were heterogeneous in the sense that they were based on two different person characteristics, which are the trait and the ERS. If the data-generating process had been a unidimensional one without any response style influence, the model would have captured this by estimating the ERS variance to be close to zero (as was shown in the first article of this thesis; see Table 3 on page 16 of the article). However, even though the response selection was thus affected by content-unrelated category preferences, trait-based responding had a considerably higher impact, which is reflected in the several times higher variance. Furthermore, the importance of the trait-based process was found to be heterogeneous across the two sub-decisions, as its loading was larger for agreement (fixed to 1.00, see Equation 2.5) compared to extreme responding (estimated to be 0.70, see parameter  $\alpha$  in Equation 2.6 and 2.7). Trait-based responding nevertheless clearly dominated all sub-decisions, which suggests that respondents endeavored to provide accurate answers.

## 4.2 Dynamic Response Process Involvement

To examine whether the involvement of trait and ERS systematically varied across items, the data was further analyzed by dynamic IRTree modeling. Two dynamic models were applied, both of which included multidimensional extreme pseudo-items and comprised the three dynamic response processes of trait-based agreement, trait-based extreme responding, and ERS-based extreme responding. One of the dynamic models constrained the item-specific loadings of the person parameters by continuous functions, whereas the second, more flexible dynamic model allowed for additional random fluctuations of the loadings. The dynamic models were compared with a model assuming a static effect of

**Table 2***Dynamic Loading Trajectories in TIMSS 2019 Data*

Loading constraint	LOO IC	Trajectory estimates		
		Process	Slope	<i>SD</i> unsyst. variation
Static	36,510	Trait agree	0	0
		Trait extreme	0	0
		ERS extreme	0	0
Continuous dynamic	36,289	Trait agree	1.80 [1.41, 2.20]	0
		Trait extreme	1.18 [0.85, 1.52]	0
		ERS extreme	0.09 [-0.20, 0.37]	0
Flexible dynamic	34,758	Trait agree	1.29 [-4.75, 6.21]	2.70 [1.45, 5.01]
		Trait extreme	0.85 [-3.66, 4.54]	2.20 [1.23, 3.97]
		ERS extreme	0.08 [-1.53, 1.68]	0.77 [0.40, 1.51]

*Note.* The slope describes the difference between the endpoint and the start value of the trajectory. 95% credible intervals in brackets.

all processes, which corresponds to the model of multidimensional extreme responding described in the previous section. The results can be found in Table 2.

Model comparisons by the LOO IC demonstrated that the dynamic models fitted the data better than the static model, indicating that dynamic heterogeneity of the response process involvement was present. To investigate the extent to which systematic changes occurred, the estimates by the continuous dynamic model were analyzed since this model was shown to provide more precise trajectory estimates than the flexible one (see simulation results of the second article on page 1370 and Table A3 in the online supplement). The estimated slopes of process trajectories revealed that the influences of both trait-based agreement and trait-based extreme responding increased throughout the questionnaire. In contrast, ERS-based extreme responding was found to have a rather constant influence. These findings suggest a warm-up effect, meaning that the respondents increasingly answered on the basis of the substantive trait as they became more familiar with the construct being measured. It is important to note, however, that this is a post-hoc interpretation and contradicts the findings of the empirical application presented in the second article of this thesis (see section 3.1), which instead suggested a decreasing trait-involvement. Furthermore, the interpretability is limited by the fact that the nine analyzed items were part of a longer survey, so additional changes in the response process involvement could have occurred across multiple scales (e.g., there may be a fatigue effect across

**Table 3**

*Suitability of Dominance and Ideal Point IRFs  
in TIMSS 2019 Data*

Trait IRF	Extreme responding	LOO IC
Dominance	Unidimensional	37,806
	Multidimensional	36,510
Ideal point	Unidimensional	43,771
	Multidimensional	40,349

the survey, but warm-up effects within individual scales).

The comparison with the flexible dynamic model showed that in addition to the systematic trajectories across items, there was further variability in the process loadings. The estimates of the flexible dynamic model revealed that the loadings of the trait scattered more strongly around the respective trajectory than the loadings of the ERS. In accordance with the findings of the application in the second article, this pattern corroborates the hypothesis that trait-based judgments generally are stronger affected by item characteristics than responding based on response styles. This interpretation is well in line with the definition of response styles as content-independent preferences, which should naturally imply a low item-specific variability. Even though this preliminary conclusion is based on few empirical data examples, it supports the construct validation of the ERS.

### 4.3 Dominance and Ideal Point Item Response Functions

A further analysis was conducted to investigate whether trait-based responding followed the dominance rationale – as assumed in the above analyses – or whether the ideal point rationale was more appropriate (i.e., whether the individual trait level per se or its distance to the item location was crucial for the response selection). Therefore, two additional models were fitted to the data, both of which assumed an ideal point IRF of the trait-based response processes. Responding based on the ERS was consistently modeled as a dominance process. The first model was a simple structure IRTree model with trait-based agreement and ERS-based extreme responding, whereas the second one included multidimensional pseudo-items of extreme responding (depending on the dominance ERS and the ideal point trait). These models were compared with the corresponding models under the assumption of trait-based responding following dominance IRFs, which are equivalent to the two models used in the first analysis. The results are given in Table 3.

The LOO model comparisons suggested that a dominance IRF better captured respond-

---

ing based on the substantive trait. Although ideal point models were often shown to be more appropriate for describing attitudinal ratings in many applications (see Tay & Ng, 2018), this does not seem to be the case for the given data. Notably, multidimensional extreme pseudo-items improved the model fit compared to unidimensional ones not only under the assumption of a dominance trait (as was already discussed in a previous section) but also for the models with ideal point IRFs. Further, the choice of the trait IRF was found to have a stronger negative effect on model fit compared to the specification of unidimensional instead of multidimensional pseudo-items, though it remains to be clarified whether this is specific to the given data set or a general phenomenon.

In summary, the analyses conducted in this chapter illustrated how IRTree modeling can be used to study heterogeneous response processes in empirical data and how the findings can provide indications of the underlying cognitive processes as well as appropriate analysis models.





## 5 General Discussion

The measurement of latent characteristics through self-reported rating items is based on the assumption that respondents rely on trait-related information when providing their answers. However, additional response processes that are unrelated to the trait of interest may likewise influence the response selection, and thus, threaten the validity of the assessment. Moreover, the involved response processes may not only be associated with different person characteristics but may exhibit additional types of heterogeneity, which can distort the trait measurement even further. In this dissertation, I addressed multidimensional item responding with the aim of improving the measurement and understanding of response process heterogeneity. I realized this research within the framework of the IRTree models, which are particularly well suited to analyze different forms of response process heterogeneity in a theory-driven way. In the three articles presented in this thesis, I evaluated the IRTree model class with regard to its capability to account for heterogeneous response processes, and I proposed further developments tailored to specific forms of heterogeneity. Thereby, I sought to contribute to the psychometric literature with a focus on two objectives: On the one hand, I investigated how modeling heterogeneous response processes can facilitate accurate measurement of latent traits. On the other hand, I explored how IRTree models can be used to gain new insights into the cognitive processes underlying item responding. The implications following from this research are discussed in the following and refer to the measurement and understanding of response process heterogeneity as well as to the IRTree model class in general.

### 5.1 Measurement and Understanding of Response Process Heterogeneity by IRTree Modeling

The response processes involved in item responding can be heterogeneous in manifold ways. The specific types of heterogeneity addressed in the three articles were the heterogeneity in terms of the person characteristics they depend on (e.g., the substantive trait or response styles), the systematically changing involvement of the processes across the items of a questionnaire (e.g., increasing impact of trait-based responding), and heterogeneity of IRFs (e.g., dominance or ideal point IRFs). Various empirical examples in the articles and the illustrative application of this thesis corroborated that these types of heterogeneity occur in real-world data and, accordingly, that disregarding such may compromise the

validity of the analysis method.

Therefore, one aim of this dissertation was to identify potential consequences of neglected response process heterogeneity and to investigate how IRTree modeling can be used to counteract such. To this end, comprehensive simulation studies were conducted to systematically evaluate the suitability of various IRTree model parameterizations under different conditions. The results demonstrated that in case the response processes indeed revealed some sort of heterogeneity, models that assumed homogeneity were at risk of providing biased estimates and misleading conclusions. In contrast, the parameter estimates of the models adapted to the specific types of heterogeneity were shown to consistently and accurately reflect the true data-generating processes. Interestingly, however, it was also found that slight model misspecifications that ignored part of the heterogeneity did not necessarily compromise the measurement of the substantive trait, which is often the primary goal of item response modeling in practice. The usefulness of different IRTree parameterizations thus depends not only on the type of response process heterogeneity present in the data but also on the objective of the application.

An important question arising from this finding is how researchers should approach the definition of IRTree models (i.e., the partitioning of ordinal responses into sub-decisions, the assignment of response processes to the pseudo-items, and the parameterization of the pseudo-items) and how they can decide on how sophisticated the model must be, or how parsimonious it can be. Traditional simple structure models, for instance, were shown to often provide acceptable trait estimates even if they were partly misspecified (e.g., if trait-based responding was ignored in sub-decisions of extreme responding). Therefore, despite the fact that the assumption of unidimensional pseudo-items does not hold in many real-world data sets (as is evident from the empirical examples in the articles as well as from other studies, e.g., Alagöz & Meiser, 2023; Jeon & De Boeck, 2016; Meiser et al., 2019), such models are probably sufficiently accurate for many practical applications, and certainly preferable to not controlling for response styles at all. On the other hand, even small errors in trait measurements can result in misleading conclusions and unfair decisions, so this risk should be minimized whenever possible. Moreover, there is neither a way to determine how severe a potential misspecification actually would be in empirical data, nor to verify that estimation biases are kept within the acceptable boundaries. Consequently, it is recommended to use unidimensional pseudo-items only if there is a theoretical or practical reason to do so and instead to consider multidimensional pseudo-items more frequently. This suggestion is particularly sensible since the erroneous inclusion of an additional response process was found to be insofar unproblematic as the estimated model parameters correctly reflected the absence of this process if it was not part of the data-generating model. Further, there exist several user-friendly R packages that allow for a straightforward implementation of multidimensional pseudo-items (e.g., see the OSF

supplement of the first article for the implementation in the *mirt* package by Chalmers, 2012), so such can be applied with little additional effort.

If researchers seek not only to improve the measurement of traits but also the understanding of the underlying processes, using even more elaborated IRTree models with a high degree of heterogeneity may be worthwhile. However, such models are more difficult to implement and may require Bayesian modeling, which limits their practical applicability. In addition, the introduction of the IRTree framework as a flexible modular system facilitates the definition of numerous different, potentially very complex models, which may provide a good fit to the data but at the cost of the interpretability of model parameters. Heavily parameterized pseudo-items should thus not be specified in a purely data-driven way, and the key criterion for defining IRTree models should generally be the theoretical foundation.

Overall, the literature suggests that IRTree modeling is appreciated and applied because of its simplicity rather than its wide-ranging scope. Therefore, the benefits of this model class could probably be better exploited if researchers made greater use of its flexibility by considering more diverse types of pseudo-item definitions than has been done so far. In this dissertation, I illustrated the advantages of such parameterizations that go beyond the traditional model specifications – in the hope that this will encourage applications and further developments of IRTree models for heterogeneous response processes.

## 5.2 Limitations and Future Directions

The articles included in this thesis integrated different types of response process heterogeneity in the IRTree framework, though there are certainly some limitations that should be addressed in future research. A main shortcoming of the conducted research is that only a subset of the hypothesized types of response process heterogeneity was covered, so it remains to be clarified to what extent the findings and modeling approaches can be generalized beyond the specific conditions: First, response style-based responding was the only type of trait-unrelated response processes considered, despite the fact that other processes may likewise affect the selection of rating categories. For example, high-stakes assessments can trigger socially desirable responding, and careless and random responding may occur in low-stakes surveys. Although such are rarely investigated by means of IRTree modeling, it may be possible to modify the models presented here to account for these additional response processes and their heterogeneity as well. However, while all trait-unrelated processes have in common that they can distort the measurement, their manifestations in the usage of the rating scale can strongly differ (e.g., socially desirable categories vary across items; careless responding can lead to various response patterns), so future research would be needed to develop an effective implementation within the IRTree

class.

A further limitation of the scope of this dissertation is that response styles were primarily incorporated by means of one specific example, the ERS. Though ERS is one of the most prevalent and most studied response styles in the literature, midscale and acquiescence response styles have been shown to frequently occur in empirical data as well (Bolt & Newton, 2011; Van Vaerenbergh & Thomas, 2013; Wetzel et al., 2016). Midscale response style is conceptually similar to ERS (they are sometimes even regarded as opposite poles of the same continuum) and can be easily included in IRTree models, as has already been illustrated in application examples in the second and the third article of this thesis. Acquiescence response style, on the other hand, is somewhat intertwined with trait-based responding in IRTree models, as both relate primarily to agreement sub-decisions. It is, therefore, more difficult to separate these two factors, especially if few reverse-keyed items are used (Park & Wu, 2019; Plieninger & Heck, 2018), so the integration of the acquiescence response style into complex models with a high degree of response process heterogeneity may be challenging.

In addition, this dissertation only targeted response process heterogeneity that is specific to the context of the measurement and can be expected to equally apply to all respondents. However, as pointed out before, it seems likely that the cognitive processing of items varies across respondents, so between-person heterogeneity should be addressed in future research. One possible approach would be the extension of the IRTree framework by mixture modeling, which has already been realized in the literature with the aim of identifying latent classes of respondents who differ in whether and which response styles they use (e.g., Alagöz & Meiser, 2023; Khorramdel et al., 2019; Kim & Bolt, 2021). Also for the dynamic IRTree models presented in the second article, it was shown that it is worth considering a person-mixture of response strategies (Merhof & Meiser, 2022). Further investigations regarding a potential mixture of dominance and ideal point responding likewise seem reasonable and may provide interesting new insights into how respondents differ in their interpretation of the items and the response scale. Besides mixture modeling, there exist alternative approaches that can even take gradual differences between respondents into account. Nevertheless, the group-specific estimation of parameters might be more feasible in practice, especially if the pseudo-items are parameterized by comparably complex IRT models.

Lastly, it should be mentioned that although response process heterogeneity can be particularly well addressed by IRTree modeling, several other approaches are available to increase the validity of trait measurements. On the one hand, various IRT models have been proposed to incorporate response styles or other trait-unrelated influences in the analysis (e.g., Falk & Cai, 2016; Henninger & Meiser, 2020; Leng et al., 2020; Scherbaum et al., 2013; Ulitzsch et al., 2022). On the other hand, there is a growing body of litera-

ture on the usage of process data collected in online assessments, such as response times, mouse movements, clickstream data, or even geolocalization and physical activity. Such data can provide an indication of how respondents engage with the items and help validate the measurement through rating scales (e.g., Keusch & Conrad, 2022; Lindner & Greiff, 2023). Further, instead of increasing the quality of analysis methods or the richness of rating data in order to improve trait measurements, other approaches were designed to prevent trait-unrelated responding in the first place. For instance, the multidimensional forced-choice format is an alternative to the rating scale and was shown to counteract socially desirable responding and response styles (Brown & Maydeu-Olivares, 2011; Wetzel et al., 2020). Other alternatives are implicit methods, such as the implicit association test (Greenwald et al., 1998), or indirect survey methods, such as randomized response techniques (e.g., Reiber et al., 2023; Warner, 1965). While these alternatives offer advantages for certain research questions and can mitigate or even prevent biases through trait-unrelated processes, the test construction, assessment, and evaluation can be considerably more complex and time-consuming compared to rating items. It is, therefore, not surprising that the vast majority of self-reported assessments relies on rating scales, and that much research is devoted to the question of how these can be constructed and analyzed in the best possible way (e.g., DeCastellarnau, 2018; Saris & Gallhofer, 2007; Wetzel & Greiff, 2018).

### 5.3 Conclusion

Rating scales are an integral part of psychology and the social sciences since they offer a straightforward method to measure characteristics of individuals that cannot be observed directly. IRTree modeling provides a tool for analyzing rating data as well as for investigating the involved response processes in a theory-driven way. In this dissertation, I demonstrated that response processes can exhibit various types of heterogeneity, and I improved the measurement and understanding of such by IRTree modeling. Although additional research is needed to better understand how heterogeneous response processes arise, what consequences they might have, and how they can be integrated into analysis models, the conducted research lays the foundation for further enhancements in the field of psychometric measurements.



## 6 Bibliography

- Alagöz, E., & Meiser, T. (2023). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, Advance online publication. <https://doi.org/10.1177/00131644231206765>
- Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 347–365. <https://doi.org/10.1111/j.2044-8317.1996.tb01093.x>
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276. <https://doi.org/10.1177/014662169301700307>
- Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. <https://doi.org/10.1037/a0028111>
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83. <https://doi.org/10.1037/met0000106>
- Böckenholt, U. (2019). Assessing item-feature effects with item response tree models. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 486–500. <https://doi.org/10.1111/bmsp.12163>
- Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. <https://doi.org/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*(5), 335–352. <https://doi.org/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*(4), 528–541. <https://doi.org/10.1037/met0000016>

- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814–833. <https://doi.org/10.1177/0013164410388411>
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. <https://doi.org/10.1177/0013164410375112>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88–106. <https://doi.org/10.1037/1040-3590.19.1.88>
- 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. <https://doi.org/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), 1–28. <https://doi.org/10.18637/jss.v048.c01>
- Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement, 54*(3), 333–363. <https://doi.org/10.1111/jedm.12147>
- DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity, 52*(4), 1523–1559. <https://doi.org/10.1007/s11135-017-0533-4>
- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3*(4), 465–476. <https://doi.org/10.1111/j.1754-9434.2010.01273.x>
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*(2), 241–261. <https://doi.org/10.1007/BF02294377>
- Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541–562. <https://doi.org/10.1037/met0000083>



- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods, 8*(1), 38–60. <https://doi.org/10.1037/1082-989X.8.1.38>
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*(3), 328–347. <https://doi.org/10.1037/met0000059>
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. <https://doi.org/10.1037/0022-3514.74.6.1464>
- Heiser, W. J. (2023). Early roots of psychometrics before Francis Galton. In L. A. van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on contemporary psychometrics* (pp. 3–30). Springer. [https://doi.org/10.1007/978-3-031-10370-4\\_1](https://doi.org/10.1007/978-3-031-10370-4_1)
- Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods, 25*(5), 560–576. <https://doi.org/10.1037/met0000249>
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. <https://doi.org/10.3758/s13428-015-0631-y>
- Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology, 72*(3), 517–537. <https://doi.org/10.1111/bmsp.12182>
- Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2022). A new multiprocess IRT model with ideal points for Likert-type items. *Journal of Educational and Behavioral Statistics, 47*(3), 297–321. <https://doi.org/10.3102/10769986211057160>
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>
- 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*(1), 92–114. <https://doi.org/10.3102/1076998609340529>
- Keusch, F., & Conrad, F. G. (2022). Using smartphones to capture and combine self-reports and passively measured behavior in social research. *Journal of Survey Statistics and Methodology, 10*(4), 863–885. <https://doi.org/10.1093/jssam/smab035>

- Khorrandel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 538–559. <https://doi.org/10.1111/bmsp.12179>
- Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. <https://doi.org/10.1177/0013164420913915>
- LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, *12*(2), 296–319. <https://doi.org/10.1177/1094428107302903>
- Leng, C.-H., Huang, H.-Y., & Yao, G. (2020). A social desirability item response theory model: Retrieve-deceive-transfer. *Psychometrika*, *85*(1), 56–74. <https://doi.org/10.1007/s11336-019-09689-y>
- Leventhal, B. C. (2019). Extreme response style: A simulation study comparison of three multidimensional item response models. *Applied Psychological Measurement*, *43*(4), 322–335. <https://doi.org/10.1177/0146621618789392>
- Lindner, M. A., & Greiff, S. (2023). Process data in computer-based assessment. *European Journal of Psychological Assessment*, *39*(4), 241–251. <https://doi.org/10.1027/1015-5759/a000790>
- Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling*, *59*(2), 183–205.
- Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*(4), 615–633. <https://doi.org/10.1007/s11336-012-9288-y>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. <https://doi.org/10.1007/BF02296272>
- McReynolds, P., & Ludwig, K. (1987). On the history of rating scales. *Personality and Individual Differences*, *8*(2), 281–283. [https://doi.org/10.1016/0191-8869\(87\)90188-7](https://doi.org/10.1016/0191-8869(87)90188-7)
- 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 and Statistical Psychology*, *72*(3), 501–516. <https://doi.org/10.1111/bmsp.12158>
- Merhof, V., & Meiser, T. (2022, July 11–15). *Non-compliant survey responding: An IRTree model for dynamically changing response strategies* [Conference presentation]. International Meeting of the Psychometric Society (IMPS), Bologna, Italy.

- Merhof, V., & Meiser, T. (2023, June 26–30). *A multidimensional IRT model of dominance and ideal point response processes* [Poster session]. SMiP Summer School, Mannheim, Germany.
- Morren, M., Gelissen, J., & Vermunt, J. (2012). The impact of controlling for extreme responding on measurement equivalence in cross-cultural research. *Methodology*, 8(4), 159–170. <https://doi.org/10.1027/1614-2241/a000048>
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- OpenAI. (2023). *ChatGPT (May 12 version)* [Large language model]. <https://chat.openai.com/chat>
- Park, M., & Wu, A. D. (2019). Item response tree models to investigate acquiescence and extreme response styles in Likert-type rating scales. *Educational and Psychological Measurement*, 79(5), 911–930. <https://doi.org/10.1177/0013164419829855>
- 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). Academic Press. <https://doi.org/10.1016/B978-0-12-590241-0.50006-X>
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun & D. N. Jackson (Eds.), *The role of constructs in psychological and educational measurement*. Routledge.
- Plieninger, H. (2020). Developing and applying IR-tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods*, 24(3), 654–670. <https://doi.org/10.1177/1094428120911096>
- Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, 53(5), 633–654. <https://doi.org/10.1080/00273171.2018.1469966>
- Ramul, K. (1963). Some early measurements and ratings in psychology. *American Psychologist*, 18(10), 653–659. <https://doi.org/10.1037/h0040858>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Reiber, F., Pope, H., & Ulrich, R. (2023). Cheater detection using the unrelated question model. *Sociological Methods and Research*, 52(1), 389–411. <https://doi.org/10.1177/0049124120914919>
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, 24(1), 3–32. <https://doi.org/10.1177/01466216000241001>

- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231–255. <https://doi.org/10.1177/014662169602000305>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*(Suppl 1), 1–97. <https://doi.org/10.1007/BF03372160>
- Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley. <https://doi.org/10.1002/9780470165195>
- Scherbaum, C. A., Sabet, J., Kern, M. J., & Agnello, P. (2013). Examining faking on personality inventories using unfolding item response theory models. *Journal of Personality Assessment, 95*(2), 207–216. <https://doi.org/10.1080/00223891.2012.725439>
- Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25–39. <https://doi.org/10.1037/0021-9010.91.1.25>
- Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods, 25*(3), 490–512. <https://doi.org/10.1177/10944281211002904>
- Tay, L., & Ng, V. (2018). Ideal point modeling of non-cognitive constructs: Review and recommendations for research. *Frontiers in Psychology, 9*, 2423. <https://doi.org/10.3389/fpsyg.2018.02423>
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics, 38*(5), 522–547. <https://doi.org/10.3102/1076998613481500>
- Thomasius, C. (1692a). *Die neue Erfindung einer wohlbegründeten und für das gemeine Wesen höchstnöthigen Wissenschaft das Verborgene des Herzens anderer Menschen auch wider ihren Willen aus der täglichen Conversation zu erkennen [New discovery of a solid science, most necessary for the community for discerning the secrets of the heart of other men from daily conversation, even against their will]*. Christoph Salfeld.
- Thomasius, C. (1692b). *Weitere Erleuterung durch unterschiedene Exempel des ohnelängst gethanen Vorschlags wegen der neuen Wissenschaft anderer Menschen Gemüther erkennen zu lernen [Further elucidation by different examples of the recent proposal for a new science for discerning the nature of other men's minds]*. Christoph Salfeld.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*(1), 39–55. <https://doi.org/10.1111/j.2044-8317.1990.tb00925.x>

- Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425–453. <https://doi.org/10.1080/00273171.2019.1643699>
- Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. <https://doi.org/10.1111/bmsp.12272>
- van Rijn, P. W., & Ali, U. S. (2018). A generalized speed-accuracy response model for dichotomous items. *Psychometrika*, *83*(1), 109–131. <https://doi.org/10.1007/s11336-017-9590-9>
- 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*(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- Verhelst, N. D., Glas, C. A. W., & de Vries, H. 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. 123–138). Springer. [https://doi.org/10.1007/978-1-4757-2691-6\\_7](https://doi.org/10.1007/978-1-4757-2691-6_7)
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69. <https://doi.org/10.1080/01621459.1965.10480775>
- Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models. *European Journal of Psychological Assessment*, *24*(1), 65–77. <https://doi.org/10.1027/1015-5759.24.1.65>
- Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, *33*(5), 352–364. <https://doi.org/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*(2), 178–189. <https://doi.org/10.1016/j.jrp.2012.10.010>
- Wetzel, E., Frick, S., & Greiff, S. (2020). The multidimensional forced-choice format as an alternative for rating scales. *European Journal of Psychological Assessment*, *36*(4), 511–515. <https://doi.org/10.1027/1015-5759/a000609>
- Wetzel, E., & Greiff, S. (2018). The world beyond rating scales. *European Journal of Psychological Assessment*, *34*(1), 1–5. <https://doi.org/10.1027/1015-5759/a000469>

- Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment, 23*(3), 279–291. <https://doi.org/10.1177/1073191115583714>
- Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*(4), 548–565. <https://doi.org/10.1177/0013164408324469>

# A Data Set, Analysis, and Model Specifications of Illustrative Application

This appendix provides further information on the illustrative application (chapter 4), including information on the empirical data set, the conducted analyses, and the IRTree models fitted to the data.

## A.1 Data Set

The data set used here is a subset of the German sample of fourth-grade students of the Trends in Mathematics and Science Study (TIMSS) from the year 2019. The analyzed scale measures the students' attitude toward learning mathematics with nine items on a four-point scale with labels "agree a lot," "agree a little," "disagree a little," and "disagree a lot." Two items (items 2 and 3) were reverse-keyed and therefore recoded. All cases with missing responses to one of the items were excluded. The raw data as well as the questionnaire can be retrieved here: <https://timss2019.org/international-database>.

From the ordinal responses, binary pseudo-item responses were derived according to the IRTree model depicted in Figure 1. The three pseudo-items are agreement ( $X_{1vi}$ ), extreme responding conditional on agreement ( $X_{2vi}$ ), and extreme responding conditional on disagreement ( $X_{3vi}$ ). The parameterization of such pseudo-items under various IRTree models is detailed below.

## A.2 Analysis

To ensure comparability of results across models introduced in different articles, the analysis scheme was standardized and may therefore differ from the ones presented in the respective articles. All models were estimated in Stan (Carpenter et al., 2017), which performs Bayesian Markov chain Monte Carlo parameter estimation. Four chains were run, each with 1,000 warmup and 1,000 post-warmup iterations. All models reached convergence, as indicated by the Gelman-Rubin statistic  $\hat{R}$  of less than 1.05.

The reported point estimates are expected a posteriori estimates. The models were

compared by out-of-sample prediction accuracy using an approximation of leave-one-out cross-validation (LOO; Vehtari et al., 2017). LOO is a fully Bayesian information criterion with good performance in IRT model selection (Luo & Al-Harbi, 2017). Small values of the LOO information criterion (LOO IC) indicate a better fit.

### A.3 Models Related to Article I

#### A.3.1 Simple Structure Model with Unidimensional Pseudo-Items

Parameterization of pseudo-items:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})} \quad (\text{A.1})$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\eta_v - \beta_{i2}))}{1 + \exp(\eta_v - \beta_{i2})} \quad (\text{A.2})$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\eta_v - \beta_{i3}))}{1 + \exp(\eta_v - \beta_{i3})} \quad (\text{A.3})$$

Priors:

$$\begin{bmatrix} \theta \\ \eta \end{bmatrix} \sim MVN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \right) \quad (\text{A.4})$$

where  $\Sigma = \text{diag}(\boldsymbol{\tau}) \times \Phi \times \text{diag}(\boldsymbol{\tau})$ ,

with  $\boldsymbol{\tau} \sim N(0, 5)$  and  $\Phi \sim LKJCorr(1)$

$$\beta \sim N(\mu_\beta, \sigma_\beta) \quad (\text{A.5})$$

with  $\mu_\beta \sim N(0, 5)$  and  $\sigma_\beta \sim N(0, 5)$

#### A.3.2 Model with Multidimensional Extreme Responding

Parameterization of pseudo-items:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})} \quad (\text{A.6})$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\eta_v + \alpha\theta_v - \beta_{i2}))}{1 + \exp(\eta_v + \alpha\theta_v - \beta_{i2})} \quad (\text{A.7})$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\eta_v - \alpha\theta_v - \beta_{i3}))}{1 + \exp(\eta_v - \alpha\theta_v - \beta_{i3})} \quad (\text{A.8})$$

with  $\alpha \geq 0$



Priors:

$$\begin{bmatrix} \theta \\ \eta \end{bmatrix} \sim MVN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{\Sigma} \right) \quad (\text{A.9})$$

where  $\mathbf{\Sigma} = \text{diag}(\boldsymbol{\tau}) \times \Phi \times \text{diag}(\boldsymbol{\tau})$ ,

with  $\boldsymbol{\tau} \sim N(0, 5)$  and  $\Phi \sim LKJCorr(1)$

$$\beta \sim N(\mu_\beta, \sigma_\beta) \quad (\text{A.10})$$

with  $\mu_\beta \sim N(0, 5)$  and  $\sigma_\beta \sim N(0, 5)$

$$\alpha \sim \text{LogN}(0, 1) \quad (\text{A.11})$$

## A.4 Models Related to Article II

### A.4.1 Continuous Dynamic Model

Parameterization of pseudo-items:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp(x_{1vi}(\alpha_{i1}^{(\theta)}\theta_v - \beta_{i1}))}{1 + \exp(\alpha_{i1}^{(\theta)}\theta_v - \beta_{i1})} \quad (\text{A.12})$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\alpha_i^{(\eta)}\eta_v + \alpha_{i2}^{(\theta)}\theta_v - \beta_{i2}))}{1 + \exp(\alpha_i^{(\eta)}\eta_v + \alpha_{i2}^{(\theta)}\theta_v - \beta_{i2})} \quad (\text{A.13})$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\alpha_i^{(\eta)}\eta_v - \alpha_{i2}^{(\theta)}\theta_v - \beta_{i3}))}{1 + \exp(\alpha_i^{(\eta)}\eta_v - \alpha_{i2}^{(\theta)}\theta_v - \beta_{i3})} \quad (\text{A.14})$$

where for each of the three processes  $p$  (trait-based agreement, trait-based extreme responding, ERS-based extreme responding), the loadings  $(\alpha_{i1}^{(\theta)}, \alpha_{i2}^{(\theta)}, \text{ and } \alpha_i^{(\eta)})$  are given by:

$$\alpha_i^{(p)} = (\gamma_1^{(p)} - \gamma_I^{(p)}) \left( 1 - \left( \frac{i-1}{I-1} \right)^{\lambda^{(p)}} \right) + \gamma_I^{(p)} \quad (\text{A.15})$$

with  $\gamma_1^{(p)} \geq 0$ ,  $\gamma_I^{(p)} \geq 0$ , and  $\lambda^{(p)} \geq 0$

Priors:

$$\theta \sim N(0, 1) \quad (\text{A.16})$$

$$\eta \sim N(0, 1) \quad (\text{A.17})$$

$$\beta \sim N(\mu_\beta, \sigma_\beta) \quad (\text{A.18})$$

with  $\mu_\beta \sim N(0, 5)$  and  $\sigma_\beta \sim N(0, 5)$

$$\gamma_1^{(p)} \sim \text{LogN}(0, 1) \quad (\text{A.19})$$

$$\gamma_I^{(p)} \sim \text{LogN}(0, 1) \quad (\text{A.20})$$

$$\lambda^{(p)} \sim \text{LogN}(-0.5, 1) \quad (\text{A.21})$$

#### A.4.2 Flexible Dynamic Model

Parameterization of pseudo-items:

See Equation A.12 - A.14. For each of the three processes  $p$ , the loadings are given by:

$$\alpha_i^{(p)} \sim \text{Normal}(\mu_i^{(p)}, \sigma^{(p)}) \quad (\text{A.22})$$

$$\mu_i^{(p)} = (\gamma_1^{(p)} - \gamma_I^{(p)}) \left( 1 - \left( \frac{i-1}{I-1} \right)^{\lambda^{(p)}} \right) + \gamma_I^{(p)} \quad (\text{A.23})$$

$$\sigma^{(p)} \sim \text{Cauchy}(0, 5) \quad (\text{A.24})$$

with  $\gamma_1^{(p)} \geq 0$ ,  $\gamma_I^{(p)} \geq 0$ , and  $\lambda^{(p)} \geq 0$

Priors:

See Equation A.16 - A.21

## A.5 Models Related to Article III

### A.5.1 Simple Structure Ideal Point Model

Parameterization of pseudo-items:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp\left(x_{1vi}(\theta_v - \delta_i) - \sum_{k=0}^{x_{1vi}} \tau_{ik}\right) + \exp\left((3 - x_{1vi})(\theta_v - \delta_i) - \sum_{k=0}^{x_{1vi}} \tau_{ik}\right)}{\sum_{j=0}^1 \left\{ \exp\left(j(\theta_v - \delta_i) - \sum_{k=0}^j \tau_{ik}\right) + \exp\left((3 - j)(\theta_v - \delta_i) - \sum_{k=0}^j \tau_{ik}\right) \right\}} \quad (\text{A.25})$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp(x_{2vi}(\eta_v - \beta_{i2}))}{1 + \exp(\eta_v - \beta_{i2})} \quad (\text{A.26})$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp(x_{3vi}(\eta_v - \beta_{i3}))}{1 + \exp(\eta_v - \beta_{i3})} \quad (\text{A.27})$$

with  $\tau_{i0} := 0$

Priors:

$$\theta \sim N(0, 1) \quad (\text{A.28})$$

$$\eta \sim N(0, 1) \quad (\text{A.29})$$

$$\beta \sim N(\mu_\beta, \sigma_\beta) \quad (\text{A.30})$$

with  $\mu_\beta \sim N(0, 5)$  and  $\sigma_\beta \sim N(0, 5)$

$$\delta \sim N(\mu_\delta, \sigma_\delta) \quad (\text{A.31})$$

with  $\mu_\delta \sim N(0, 5)$  and  $\sigma_\delta \sim N(0, 5)$

$$\tau \sim N(0, 5) \quad (\text{A.32})$$

### A.5.2 Ideal Point Model with Multidimensional Extreme Responding

Parameterization of pseudo-items:

$$p(X_{1vi} = x_{1vi}) = \frac{\exp\left(x_{1vi}(\theta_v - \delta_i) - \sum_{k=0}^{x_{1vi}} \tau_{1ik}\right) + \exp\left((3 - x_{1vi})(\theta_v - \delta_i) - \sum_{k=0}^{x_{1vi}} \tau_{1ik}\right)}{\sum_{j=0}^1 \left\{ \exp\left(j(\theta_v - \delta_i) - \sum_{k=0}^j \tau_{1ik}\right) + \exp\left((3 - j)(\theta_v - \delta_i) - \sum_{k=0}^j \tau_{1ik}\right) \right\}} \quad (\text{A.33})$$

$$p(X_{2vi} = x_{2vi}) = \frac{\exp\left(x_{2vi} \alpha(\theta_v - \delta_i) + x_{2vi} \eta - \sum_{k=0}^{x_{2vi}} \tau_{2ik}\right) + \exp\left((3 - x_{2vi}) \alpha(\theta_v - \delta_i) + x_{2vi} \eta - \sum_{k=0}^{x_{2vi}} \tau_{2ik}\right)}{\sum_{j=0}^1 \left\{ \exp\left(j \alpha(\theta_v - \delta_i) + j \eta - \sum_{k=0}^j \tau_{2ik}\right) + \exp\left((3 - j) \alpha(\theta_v - \delta_i) + j \eta - \sum_{k=0}^j \tau_{2ik}\right) \right\}} \quad (\text{A.34})$$

$$p(X_{3vi} = x_{3vi}) = \frac{\exp\left((1 - x_{3vi}) \alpha(\theta_v - \delta_i) + x_{3vi} \eta - \sum_{k=0}^{x_{3vi}} \tau_{3ik}\right) + \exp\left((2 + x_{3vi}) \alpha(\theta_v - \delta_i) + x_{3vi} \eta - \sum_{k=0}^{x_{3vi}} \tau_{3ik}\right)}{\sum_{j=0}^1 \left\{ \exp\left((1 - j) \alpha(\theta_v - \delta_i) + j \eta - \sum_{k=0}^j \tau_{3ik}\right) + \exp\left((2 + j) \alpha(\theta_v - \delta_i) + j \eta - \sum_{k=0}^j \tau_{3ik}\right) \right\}} \quad (\text{A.35})$$

with  $\tau_{1i0} := 0$ ,  $\tau_{2i0} := 0$ , and  $\tau_{3i0} := 0$

Priors:

$$\theta \sim N(0, 1) \quad (\text{A.36})$$

$$\eta \sim N(0, 1) \quad (\text{A.37})$$

$$\delta \sim N(\mu_\delta, \sigma_\delta) \quad (\text{A.38})$$

with  $\mu_\delta \sim N(0, 5)$  and  $\sigma_\delta \sim N(0, 5)$

$$\tau \sim N(0, 5) \quad (\text{A.39})$$

$$\alpha \sim \text{LogN}(0, 1) \quad (\text{A.40})$$

## **B Copies of Articles**



# Separation of Traits and Extreme Response Style in IRTree Models: The Role of Mimicry Effects for the Meaningful Interpretation of Estimates

Educational and Psychological  
Measurement  
1–30

© The Author(s) 2023



Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/00131644231213319  
journals.sagepub.com/home/epm



Viola Merhof<sup>1</sup> , Caroline M. Böhm<sup>2</sup> and Thorsten Meiser<sup>1</sup>

## Abstract

Item response tree (IRTree) models are a flexible framework to control self-reported trait measurements for response styles. To this end, IRTree models decompose the responses to rating items into sub-decisions, which are assumed to be made on the basis of either the trait being measured or a response style, whereby the effects of such person parameters can be separated from each other. Here we investigate conditions under which the substantive meanings of estimated extreme response style parameters are potentially invalid and do not correspond to the meanings attributed to them, that is, content-unrelated category preferences. Rather, the response style factor may mimic the trait and capture part of the trait-induced variance in item responding, thus impairing the meaningful separation of the person parameters. Such a mimicry effect is manifested in a biased estimation of the covariance of response style and trait, as well as in an overestimation of the response style variance. Both can lead to severely misleading conclusions drawn from IRTree analyses. A series of simulation studies reveals that mimicry effects depend on the distribution of observed responses and that the estimation biases are stronger the more asymmetrically the responses are distributed across the rating scale. It is further demonstrated that extending the commonly used IRTree model with unidimensional sub-decisions by multidimensional parameterizations counteracts mimicry effects and facilitates the meaningful separation of parameters. An empirical example of the Program for

<sup>1</sup>University of Mannheim, Germany

<sup>2</sup>Rhineland-Palatinate Technical University of Kaiserslautern-Landau, Germany

## Corresponding Author:

Viola Merhof, Department of Psychology, University of Mannheim, L 13, 15, D-68161 Mannheim, Germany.  
Email: merhof@uni-mannheim.de

International Student Assessment (PISA) background questionnaire illustrates the threat of mimicry effects in real data. The implications of applying IRTree models for empirical research questions are discussed.

### **Keywords**

IRTree models, response styles, multidimensional item responding, meaningful model parameters

Item response tree (IRTree) models are a popular class of multidimensional item response theory (IRT) approaches for analyzing self-reported Likert-type rating data (Böckenholt, 2012; De Boeck & Partchev, 2012). They rest on the assumption that item responding comprises several qualitatively distinct judgment steps, which are processed by respondents on the basis of different latent personal traits. A typical aim of using IRTree models is to separate the effects of the substantive trait to be measured from those of response styles, which are individual preferences for specific response categories of rating scales irrespective of item content (for an overview, see Van Vaerenbergh & Thomas, 2013). For instance, respondents may prefer extreme over non-extreme categories (extreme response style; ERS) or they tend to choose the middle categories of a scale (midscale response style; MRS). Since such different ways of using the rating scale can systematically bias trait estimates, there is great interest in both research and practice to apply methods that account for response styles and thereby provide valid trait measurements (Baumgartner & Steenkamp, 2001).

IRTree models provide an easy-to-implement framework for specifying various response styles in a theory-driven way. The ordinal responses to rating items are split into meaningful sub-decisions, which are modeled to be made on the basis of either the content-related trait or a response style. For example, respondents may first take a trait-based decision on whether they generally agree or disagree with the item, and subsequently select one of the available categories reflecting more or less intense agreement or disagreement driven by their response styles. Such sub-decisions are typically parameterized by unidimensional IRT models (e.g., the Rasch or 2PL model), so that the multidimensionality of IRTree models arises only between the sub-decisions, thus keeping the modeling complexity low and providing a straightforward interpretation of the parameters.

Several studies have demonstrated that IRTree models successfully capture multidimensional item responding, and such models were used for controlling trait measurements for response styles in various applications (e.g., Böckenholt & Meiser, 2017; Jeon & De Boeck, 2016; Khorramdel & von Davier, 2014; Kim & Bolt, 2021; Plieninger & Meiser, 2014; Tijmstra et al., 2018). However, the previous research solely focused on the assumption that response styles were actually involved in the item response process, so it is unclear how IRTree models perform in the absence of



any response style effect. Even though the assignment of response styles to certain sub-decisions is theoretically founded, there may be circumstances in which respondents nevertheless make all their judgments solely on the basis of the trait; for example, if the respondents have a great interest in providing accurate information like in high-stakes assessments (e.g., personality assessments in job interviews). Since IRTree models are over-parameterized in such cases (i.e., they include several person parameters for modeling unidimensional data), they might be prone to overfitting and could carry the risk of estimating response style variance which is non-existent. Therefore, it remains to be investigated under which conditions the estimates obtained by IRTree modeling successfully reflect the substantive meaning assigned to the parameters and under which conditions they do not. Answering this question is of high relevance, as the potential lack of validity may compromise the key characteristic of IRTree models, which is their ability to disentangle the influences of multiple person parameters.

In addition, it is of particular concern that the estimated parameters labeled as response styles in misspecified IRTree models not only absorb random variance but may rather reflect trait-based responding and capture variance induced by the substantive trait. Since the response style factor then would mimic part of the trait, we call this methodological artifact a *mimicry effect*. The occurrence of such implies that the separability of traits and response styles is compromised, that is, the variance components in item responding are partially misattributed to a factor that is not the true source of the variance. Therefore, the mimicry effect is primarily manifested in a biased estimation of the relationship between trait and response style (i.e., their covariance and correlation). Furthermore, the variance of the response style factor is overestimated, as it captures additional trait-induced variance.

As a result, the pitfalls of mimicry effects in IRTree models for drawing conclusions from the data are twofold: First, the influences of content-unrelated category preferences are overestimated. Accordingly, even the dimensionality of the response process might be overestimated by an IRTree model if a response style was estimated to vary across respondents, despite not being part of the actual data-generating process. Second, the substantive meaning of the response style factor no longer corresponds to the meaning that was assigned to it, as the estimates at least partly reflect trait-based responding. Although the meanings of response styles and traits then overlap, they are considered distinct response processes given their associations with qualitatively different sub-decisions. Moreover, one might even find reasonable theoretical justifications for correlations between traits and given response styles post hoc (e.g., respondents with high levels of extraversion are likely to favor extreme response categories because they are generally self-confident), so that no further attention would be paid to an artificially induced covariance.

A likely scenario for an impaired separability of person parameters by IRTree models arises when the distribution of the respondents' trait levels differs from that of the items of the questionnaire, such as when the trait follows a skewed or shifted distribution. For instance, skewed or shifted distributions are to be expected if the

questionnaire was originally generated for a different sub-population of respondents with substantially higher or lower trait levels, or if a very rare or common trait is being assessed. An example would be questionnaires designed to measure the severity of mental disorders, for which the majority of the population scores low (e.g., the Beck Depression Inventory; Beck et al., 1996). Furthermore, many scales developed for the assessment of personality or attitudes of the general population have expected scores above the scale mean (e.g., five-factor models of personality like the International Personality Item Pool scales; Goldberg et al., 2006), while other scales have expected scores below the scale mean (e.g., the Dark Factor of Personality; Moshagen et al., 2018). All of these response patterns are likely the result of either skewed or shifted trait distributions in relation to the respective item distributions. Although it is thus inherently clear for a variety of empirical research questions that the distributions mismatch, there has been no systematic investigation of how this affects the parameter estimation and validity of IRTree models.

Therefore, the aim of this article is to evaluate the parameter estimation of IRTree models for various trait distributions with a focus on the separability of person parameters. Thereby, this article is intended to increase awareness for potentially biased estimates of response style parameters, in which case their assigned substantive meanings are invalid. Conversely, this does not mean that if person parameters are successfully separated by a model and statistically unbiased estimates are obtained, these estimates actually reflect the attributed substantive meaning in the sense of content validity. Yet, there is some evidence in the literature in favor of the validity of response style estimates: Investigations of the criterion validity of response styles showed that the IRTree estimates were linked to extraneous criteria as one would theoretically expect (Plieninger & Meiser, 2014; Zhang & Wang, 2020). In addition, individual response style estimates were found to be stable across different constructs (Wetzel et al., 2013) and over time (Weijters et al., 2010; Wetzel et al., 2016), which does not provide evidence for the validity per se, but still suggests that response styles are trait-like constructs and a characteristic of the persons rather than of the items or questionnaires. In a combined analysis of rating responses and response times, it was further revealed that responses that matched the person-specific response styles were faster, as one would expect given the conception of response styles as heuristic response processes (Henninger & Plieninger, 2020). Although these results support the use of IRT models accommodating response styles, such as IRTree models, the substantive validity of estimates can never be achieved without the accurate separation of traits and response styles. Therefore, with the analysis of the parameter separation in the present study, we are laying the groundwork for further investigations of the validity of IRTree models.

In the next section, IRTree models are formally introduced and the challenge of a meaningful separation of response style parameters from substantive traits is illustrated. Then, a series of three simulation studies is presented that examine the conditions under which IRTree models are at risk of compromised separability. Thereby, we quantify mimicry effects and explore how such a potential lack of validity can be

detected. Since the main purpose of response style modeling in empirical research and practice is to obtain unbiased trait measurements, we additionally investigate the impact of mimicry effects on the recovery of the person-specific trait levels. In Simulation Study 1, the extent to which mimicry effects occur depending on the distribution of the substantive trait in relation to the items is assessed. In Simulation Study 2, potential remedies are evaluated with respect to their capability to counteract mimicry effects. As these two studies focus on the potential risk of applying IRTree models to data originating from a unidimensional response process (i.e., where the respondents' decisions are purely trait-based), we investigate whether the findings are transferable to multidimensional data with the combined influences of trait and response styles in Simulation Study 3. Thereafter, an empirical application to the background questionnaire of the Program for International Student Assessment (PISA) 2018 study is presented, which demonstrates the threat of mimicry effects in real data. Finally, the results are discussed and implications for using IRTree models in empirical research are derived.

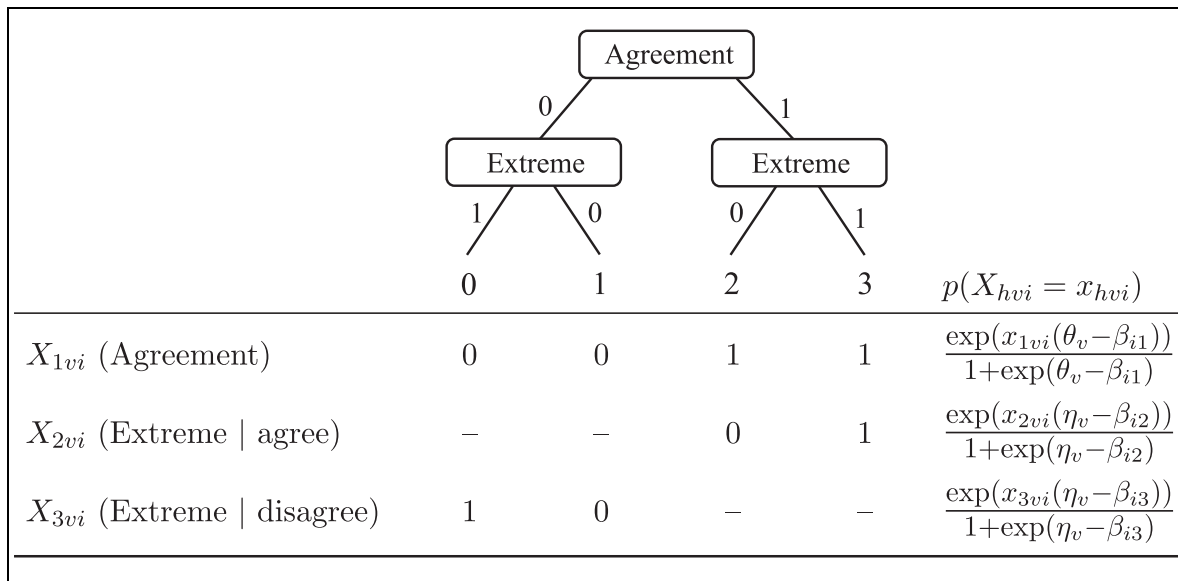
## Separation of Traits and Response Styles in IRTree Models

IRTree models decompose the ordinal rating responses  $Y_{vi} \in \{0, \dots, K\}$  of person  $v = 1, \dots, N$  to item  $i = 1, \dots, I$  into a sequence of binary pseudo-items  $X_{hvi}$ , which represent the sub-decisions assumed to be taken by respondents during the response selection. The pseudo-items are usually parameterized by unidimensional IRT models of the trait or a response style, and the probability of an ordinal response is the product of the probabilities of responses to the respective pseudo-items. Figure 1 depicts a commonly used two-dimensional IRTree model for responding to items on a 4-point scale, with one sub-decision reflecting trait-based agreement, and a second one ERS-based extreme responding conditional on agreement. The pseudo-item responses are parameterized by Rasch models of either the substantive trait  $\theta$  ( $h = 1$ ) or the ERS  $\eta$  ( $h = 2$  and  $h = 3$ ), and the probability of an ordinal response  $Y_{vi} \in \{0, \dots, 3\}$  is obtained by

$$p(Y_{vi}=y_{vi}) = \left[ \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})} \right] \times \left[ \frac{\exp(x_{2vi}(\eta_v - \beta_{i2}))}{1 + \exp(\eta_v - \beta_{i2})} \right]^{x_{1vi}} \times \left[ \frac{\exp(x_{3vi}(\eta_v - \beta_{i3}))}{1 + \exp(\eta_v - \beta_{i3})} \right]^{(1-x_{1vi})}, \quad (1)$$

where  $\beta_{ih}$  denotes the difficulty of pseudo-item  $h$  of item  $i$ .

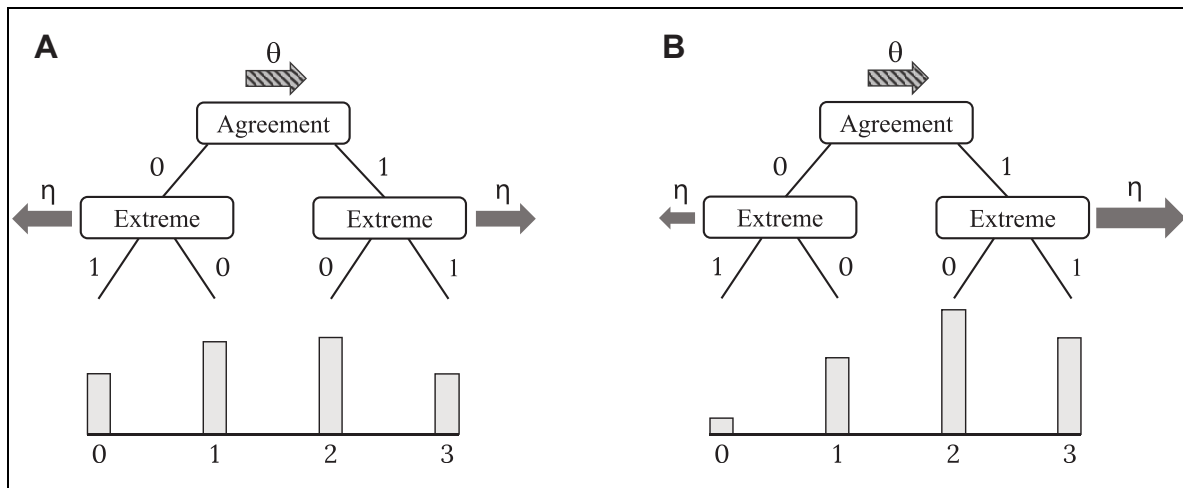
The separation of traits and response styles in IRTree models is achieved by defining model structures in which (a) the different personal characteristics are related to different pseudo-items (e.g., trait-based agreement and ERS-based extreme responding) and (b) they affect the selection of ordinal categories in unique ways that cannot be linearly transformed into each other (e.g., high trait levels favor high categories and high ERS levels favor extreme/outer categories). The first property leads to the identification of the model and enables the estimation of several parameters for each



**Figure 1.** Tree Diagram and Definition of Pseudo-Items for Responses to 4-Point Rating Items. Note. Pseudo-items missing by design are marked with “–”.

respondent. As these person parameters are assigned to different pseudo-items, a non-redundant part of the information from the ordinal responses is available for the estimation of each of them, and they can be statistically separated from each other. Nonetheless, only the second property ensures a meaningful distinction and substantive separation of different person parameters. Figure 2 illustrates the uniquely directed influences of trait and ERS for the IRTree model of 4-point rating items: Higher substantive trait levels are modeled to increase the probability of selecting agreement (i.e., higher) categories, whereas higher ERS levels favor the extreme categories, and thus either affect the response selection in the direction of the trait (i.e., higher categories conditional on agreement) or in the opposite direction (i.e., lower categories conditional on disagreement). The ERS factor is therefore assigned the substantive meaning of a preference for extreme categories based on its effects on the category selection across the two pseudo-items. It can only capture variance in the respondents' behavior which equally affects the choice of extreme agreement and extreme disagreement categories.

Although such a definition of substantive traits and response styles as unique influences on different sub-decisions is theoretically reasonable, IRTree models may be misspecified such that the importance of a response style is being overstated or, correspondingly, the influence of the trait is being understated by the model. For instance, the true data-generating process could be a unidimensional one without any response style influence on the judgment process. Fitting an IRTree model assuming a response style influence (like the one depicted in Figure 1) to such unidimensional data entails the risk of a mimicry effect, as the response style parameters are not required to account for individual category preferences, and thus may be redeclared to capture variance which was actually introduced by the substantive trait. Such a redeclaration of the response style factor in terms of taking over part of the



**Figure 2.** Illustration of the Meaningful Separation of Trait and ERS in an IRTree Model Depending on the Response Distribution. (A) Symmetrical Response Distribution. (B) Asymmetrical Response Distribution.

substantive trait would inevitably impede the meaningful separation of the two person parameters, though without affecting the statistical separation.

In the following, our hypotheses regarding the conditions of such an impairment of the meaningful separation of traits and response styles in IRTree models and the occurrence of mimicry effects are derived: Primarily, we expected that the statistical advantage of using response style parameters as a substitute for the substantive trait should depend on the distribution of the ordinal responses across the rating scale, which in turn is determined by the distribution of the substantive trait levels in relation to that of the items. If trait and item distributions match, the distribution of observed item responses is symmetrical (i.e., similar frequencies of agreement and disagreement categories; see Figure 2A), and a mimicry effect should not occur. In the exemplary 4-point IRTree model, the more the ERS factor would mimic the substantive trait, the better the variance among the agreement item responses should be accounted for, but the worse the variance among the disagreement categories. Thus, the congruent and opposing effects of trait and ERS should cancel out, they should have unique influences on the selection of ordinal categories, and their meaningful separation should remain intact.

In contrast, if the distribution of observed responses is asymmetrical (i.e., unequal distribution across agreement and disagreement categories; see Figure 2B), the congruent and opposing effects of trait and ERS can be assumed to not cancel out, and their influences on the selection of ordinal categories should partly overlap. A large proportion of the data should be better explained by ERS parameters mimicking the trait, and only a small proportion should be less well explained. Therefore, the model should benefit from a redeclaration of the ERS factor as a substitute for the trait, resulting in a substantial variance of the estimated ERS levels and covariance with the trait levels. Such a mimicry effect would be accompanied by a reduction or even complete loss of the meaningful separation of trait and response style, as the

trait-induced variance would be jointly explained by both parameters. The severity of mimicry effects (i.e., the degree of reduction in separability) was thus assumed to be larger, the more asymmetrical the response distribution is, and accordingly, the more skewed or shifted the trait distribution in relation to the item distribution. Thereby, the reduced separation of trait and ERS was expected to occur for both an asymmetry of the response distribution toward high or low categories of the scale. However, an asymmetrical distribution toward higher categories should be associated with a positive covariance of ERS and trait, whereas a shift toward lower categories with a negative covariance. Since a negative covariance of two parameters implies the same degree of shared meaning as a corresponding positive one, the mimicry effect should be independent of the direction of the response distribution asymmetry and quantified by the absolute covariance.

In the next section, we illustrate our expectations regarding mimicry effects and their dependency on the distribution of rating responses in a first simulation study.

### Simulation Study I—Mimicry Effects and Trait Distributions

The first simulation study addresses mimicry effects in IRTree models for various distributional conditions of the trait in the absence of response style influences. Therefore, unidimensional item response data were generated under the partial credit model (PCM; Masters, 1982), which was chosen as the data-generating model as it is a commonly used IRT model for ordinal rating data. The two-dimensional IRTree model with trait and ERS influences illustrated in Figure 1 was used as the analysis model. The underlying trait distributions were set to be either skewed or shifted in relation to the items, which both result in asymmetrical response distributions. Besides the investigation of mimicry effects, we evaluated the recovery of the substantive trait levels, which provides an indication of whether IRTree models produce reasonable estimates for individual parameters despite the risk of a biased estimation of the covariance structure.<sup>1</sup>

#### Simulation Design

Item response data were generated under the PCM, a unidimensional IRT model in which the selection of all ordinal response categories is assumed to depend solely on the substantive trait. The category probabilities of the ordinal responses  $Y_{vi} \in \{0, \dots, K\}$  under the PCM are given by

$$p(Y_{vi} = y_{vi}) = \frac{\exp\left(y\theta_v - \sum_{k=0}^y \beta_{ik}\right)}{\sum_{j=0}^K \exp\left(j\theta_v - \sum_{k=0}^j \beta_{ik}\right)}, \quad (2)$$

with  $\beta_{i0} := 0$ .  $\theta_v$  denotes the person-specific trait level and  $\beta_{ik}$  denotes the item- and category-specific difficulty. The difficulty parameters can be decomposed as

$\beta_{ik} = \beta_i + \tau_{ik}$ , where  $\beta_i$  denotes the item location and is defined as  $\sum_{k=1}^K \beta_{ik}/K$  and  $\tau_{ik}$  denotes the category-specific deviations with  $\sum_{k=1}^K \tau_{ik} = 0$ .

Responses to 4-point rating items were generated, with the category-specific difficulty parameters  $\beta_{ik}$  for each item drawn from a uniform distribution  $U(-3, 3)$  and assigned to the ordinal categories  $k = \{1, 2, 3\}$  in ascending order ( $\beta_{i0}$  is defined to be 0 in the PCM). This sampling procedure results in the item locations  $\beta_i$  being approximately normally distributed with mean 0 and variance 1.

For conditions of shifted trait distributions, person-specific trait levels  $\theta_v$  were sampled from normal distributions  $N(\mu, 1.0)$  with mean  $\mu$  set to 0.0, 0.2, 0.5, or 1.0. Therefore, the distributions of the traits and item locations either matched ( $\mu = 0.0$ ; baseline condition) or were shifted by 0.2, 0.5, or 1.0 units of the *SD*. The stronger the shifts of the trait distributions, the more the distributions of the ordinal item responses were asymmetrical toward the higher categories of the scale.

For the skewed conditions, the substantive trait was assumed to stem from a skew-normal distribution with the probability density function:

$$SkewN(x | \xi, \omega, \alpha) = \frac{2}{\omega} \phi\left(\frac{x - \xi}{\omega}\right) \Phi\left(\alpha \left(\frac{x - \xi}{\omega}\right)\right), \quad (3)$$

where  $\phi$  denotes the standard normal probability density function and  $\Phi$  denotes the cumulative distribution function (for further details on the skew-normal distribution, see Azzalini & Capitanio, 2014). The parameter  $\xi$  is the location,  $\omega$  is the scale, and  $\alpha$  is the skewness of the distribution. Positive  $\alpha$  values result in right-skewed distributions and negative values in left-skewed ones. The skew-normal distribution reduces to the standard normal one for  $\xi = 0$ ,  $\omega = 1$ , and  $\alpha = 0$ . The mean and variance of a skew-normally distributed variable  $X \sim SkewN(\xi, \omega, \alpha)$  are defined as

$$M = \xi + \frac{\omega \alpha \sqrt{2}}{\sqrt{(1 + \alpha^2)\pi}}, \quad (4)$$

$$Var = \omega^2 \left(1 - \frac{2\alpha^2}{\pi(1 + \alpha^2)}\right). \quad (5)$$

The trait levels were sampled from  $SkewN(\xi = 0, \omega, \alpha)$ , with skewness parameter  $\alpha$  set to 0.0, 0.5, 1.0, and 2.0. The corresponding scale parameters  $\omega$  were set to 1.00, 1.07, 1.21, and 1.43, resulting in the trait distributions of all conditions having a variance of 1, which provided a high degree of comparability between all shifted and skewed conditions. According to Equation 4, the means of the four conditions were 0.00, 0.38, 0.68, and 1.02. The baseline condition with skewness parameter  $\alpha = 0.0$  was equivalent to sampling from a standard normal distribution and thus equivalent to the baseline condition of the shifted data generation. The higher the skewness parameters were, the stronger the asymmetry of ordinal item responses toward the higher categories of the scale.

For each of the four shifted and four skewed conditions, random data sets were generated with varying sample sizes  $N$ , set to 100, 500, and 2,000, and questionnaire lengths  $I$ , set to 5, 10, 20, and 40. The simulation factors were fully crossed, resulting in 96 ( $8 \times 3 \times 4$ ) simulation conditions, for which 100 replications were conducted each. For each data set, item responses were generated as follows:<sup>2</sup> $N$  trait levels and  $3 \times I$  item difficulties were randomly drawn according to the sampling procedure of the respective simulation condition described above. Then, person and item parameters were inserted into the PCM given by Equation 2, yielding category-specific probabilities for responses of each person to each item. Finally, ordinal item responses were sampled according to the model-implied probabilities, and pseudo-item responses were derived from such ordinal responses according to the definition given in Figure 1.

The generated data sets were analyzed within the IRTree framework, and the model described in Figure 1 and Equation 1 was applied. We additionally used the PCM as an analysis model to obtain benchmarks for the trait recovery. The models were estimated using the R package *mirt*<sup>3</sup> (Chalmers, 2012) and all models converged.

The recovery of substantive traits was assessed by the correlation of generated and estimated (expected a posteriori) parameters. This measure of recovery was chosen as it indicates whether the ranking of the persons was correctly reflected by the model estimates. Other commonly used measures of recovery, such as the mean absolute bias, were not suitable here as the conditions with shifted or skewed trait distributions necessarily result in larger absolute deviations of estimated parameters. In addition, the rank order is a crucial measure when assessments are used as the basis for decisions in a practical context, such as the selection of the best applicants in a job interview.

Further note that for the data generation under both shifted and skewed trait distributions, only conditions resulting in an asymmetrical response distribution toward the higher categories were defined. We chose such conditions since asymmetrical distributions in the opposite direction toward low categories can be expected to not affect the size of the mimicry effects in IRTree models which have a symmetrical tree structure. The IRTree model used here has such a symmetrical structure, as the agreement sub-decision splits the rating scale into two categories each, which are then again split by the extreme sub-decisions. Thus, we assumed that the only difference in the mimicry effects for asymmetrical response distributions toward high and low categories should be that the deviation of the estimated covariance from the true covariance reverses in sign. We nevertheless run the same simulation study as described here just with reversed trait shifts ( $\mu = -0.2, -0.5, \text{ or } -1.0$ ) and reversed skewness parameters ( $\alpha = -0.5, -1.0, \text{ and } -2.0$ ). The results can be found in the Online Supplementary Materials and confirm our assumption that the corresponding positive and negative parameters resulted in nearly equivalent sizes of the mimicry effects. Therefore, only conditions associated with an asymmetry to the high categories are presented in the following.



## Results

The mimicry effects were evaluated by the estimated covariances between ERS and substantive trait levels. As the data-generating process was unidimensional and did not incorporate response style influences, the accurate estimate for the covariance would be zero. Thus, the more an estimated covariance deviated from zero, the stronger the mimicry effect was. In line with our expectations, the mimicry effects were more pronounced the more skewed or shifted the underlying trait distributions were in relation to the distribution of item locations (i.e., the more asymmetrical the item response distributions), as can be seen in Table 1. Likewise, the estimated correlations between ERS and trait strongly increased with increasingly asymmetric response distributions. For the baseline conditions with non-shifted or non-skewed distributions, the covariances and correlations were correctly estimated to be close to zero.

The conditions with shifted and skewed trait distributions hardly differed with respect to revealing increasing mimicry effects for increasing deviations from the standard normal baseline condition. However, the skewed conditions generally yielded slightly stronger effects, suggesting that the asymmetry of item responses was higher for the specific skewness parameters  $\alpha$ , compared with the specific mean shifts  $\mu$  we defined. Sample size and questionnaire length did not influence the mimicry effect, as there were only small differences in the estimated covariances across these simulation factors (see Table A1 in the Online Supplementary Materials).

Furthermore, the estimated variances of the ERS factor increased with higher shifts or skewness parameters of the trait distributions, which is in line with the assumption that the ERS parameters increasingly take over trait-induced variance for a stronger asymmetry of observed item responses. Also for the baseline conditions without mimicry effects, the ERS variances were greater than zero, indicating that

**Table 1.** Estimated Covariances, Correlations, and Variances of ERS  $\eta$  and Trait  $\theta$  by the IRTree Model for Unidimensional Data (Simulation 1).

Trait distribution		<i>M</i> ( <i>SD</i> ) across replications			
Distr. family	Condition	$\widehat{\text{Cov}}(\eta, \theta)$	$\widehat{\text{Cor}}(\eta, \theta)$	$\widehat{\text{Var}}(\eta)$	$\widehat{\text{Var}}(\theta)$
Shifted	$\mu = 0.0$	0.00 (0.26)	0.00 (0.34)	0.26 (0.13)	2.11 (0.37)
	$\mu = 0.2$	0.15 (0.26)	0.20 (0.34)	0.26 (0.12)	2.11 (0.37)
	$\mu = 0.5$	0.35 (0.27)	0.44 (0.29)	0.31 (0.15)	2.14 (0.36)
	$\mu = 1.0$	0.70 (0.26)	0.74 (0.18)	0.40 (0.16)	2.29 (0.43)
Skewed	$\alpha = 0.0$	0.01 (0.26)	0.01 (0.34)	0.26 (0.14)	2.12 (0.39)
	$\alpha = 0.5$	0.28 (0.26)	0.36 (0.30)	0.28 (0.13)	2.15 (0.38)
	$\alpha = 1.0$	0.54 (0.27)	0.62 (0.25)	0.35 (0.15)	2.18 (0.39)
	$\alpha = 2.0$	0.84 (0.23)	0.83 (0.13)	0.49 (0.18)	2.14 (0.41)

Note. Aggregated across sample sizes ( $N = 100, 500, 2,000$ ) and questionnaire lengths ( $l = 5, 10, 20, 40$ ).  
ERS = extreme response style.

the parameters still captured some variation in the selection of extreme categories across respondents. Nevertheless, even in conditions with strong mimicry effects, the average variance estimates of the ERS factor were rather small in comparison to those of the trait, which is due to the fact that the data-generating process was purely trait-based and did not include ERS-based responding. Furthermore, the estimated trait variances were much higher than the generated variances of 1, which is due to differences in the model specifications between the data-generating PCM and the analysis IRTree model.

Overall, the analysis of the variance and covariance estimation clearly demonstrated that IRTree models pose the risk of mimicry effects, and thus potentially lead to inaccurate conclusions about the involved response processes and the relationship of the person parameters. Even for trait distributions with slight shifts or small degrees of skewness, the estimated covariances of trait and ERS were of substantial size. Such estimates would likely lead researchers to the erroneous interpretation that respondents with high levels of the substantive trait had strong preferences for extreme categories and those with low trait levels rather preferred the non-extreme ones, when in fact, the respondents did not at all have category preferences.

However, since a main use case of IRTree modeling is to obtain accurate trait measurements that are controlled for response style influences, the response style estimates themselves or covariances with other parameters are often not of interest. Therefore, we additionally examined the recovery of the substantive trait levels. Notably, the presence of a mimicry effect did not impair the trait recovery by the IRTree model. Irrespective of the distributional condition, the correlations of generated and estimated trait levels were consistently high, as is evident from Table 2 (also see Table A2 in the Online Supplementary Materials for the trait recovery split by  $N$  and  $I$ ). The PCM yielded a slightly higher trait recovery in all conditions, which can be considered the benchmark or maximal achievable values of recovery,

**Table 2.** Trait Recovery  $Cor(\theta, \hat{\theta})$  by the IRTree Model and Data-Generating PCM for Unidimensional Data (Simulation 1).

Trait distribution		$M$ ( $SD$ ) across replications	
Distr. family	Condition	IRTree	PCM
Shifted	$\mu = 0.0$	0.87 (0.08)	.91 (.06)
	$\mu = 0.2$	0.87 (0.08)	.91 (.06)
	$\mu = 0.5$	0.88 (0.08)	.91 (.06)
	$\mu = 1.0$	0.88 (0.07)	.91 (.06)
Skewed	$\alpha = 0.0$	0.88 (0.07)	.91 (.06)
	$\alpha = 0.5$	0.88 (0.07)	.91 (.06)
	$\alpha = 1.0$	0.88 (0.07)	.91 (.06)
	$\alpha = 2.0$	0.88 (0.08)	.90 (.06)

Note. Aggregated across sample sizes ( $N = 100, 500, 2,000$ ) and questionnaire lengths ( $I = 5, 10, 20, 40$ ).  
 $SD$  = standard deviation; PCM = partial credit model.

as the PCM was the true data-generating model. However, the PCM uses the information of all four ordinal response categories for the estimation of the trait levels, while the IRTree model only uses the information provided by the binary agreement sub-decision, so this small advantage of the PCM is not surprising.

Thus, the potential occurrence of a mimicry effect in IRTree modeling is primarily a concern for estimating response styles, but less so for recovering person-specific trait levels. If the focus of the analysis is exclusively on the measurement of substantive traits, our results suggest that skewed or shifted trait distributions do not have relevant effects. Nonetheless, a bias in the latent covariance matrix may lead to misinterpretations regarding the item response process and involved person parameters. Therefore, we explore possible modifications of the previously used IRTree model in the next section, which potentially could counteract mimicry effects and provide unbiased estimates of all model parameters.

## **Simulation Study 2—Modified IRTree Models Counteracting Mimicry Effects**

In the second simulation study, two modified IRTree models were examined with regard to their ability to counteract mimicry effects. The first modified IRTree model differed from the standard IRTree model described before (see Figure 1 and Equation 1) in that the covariance of trait and ERS was fixed to zero. This model constraint prevents the estimation of artificial covariances, as evoked by the ERS parameters mimicking the substantive traits. Consequently, erroneous conclusions about the relationship between the trait and response styles cannot arise even in the presence of asymmetrical response distributions. However, this comes with the disadvantage that such a model cannot capture a true covariance if it would actually be present in the data, and that a zero covariance of personal characteristics is often not reasonable from a theoretical point of view.

Therefore, another modified IRTree model with freely estimated covariance was evaluated, in which the extreme pseudo-items were parameterized by multidimensional IRT models (see Böckenholt, 2019; Jeon & De Boeck, 2016; Meiser et al., 2019). Such a multidimensionality within pseudo-items (additionally to the usual multidimensionality between pseudo-items) reflects the assumption that not only one, but several person parameters are involved in the respective sub-decisions. For instance, in the IRTree model for 4-point rating items, respondents may use both the ERS and the trait for the sub-decisions of extreme versus non-extreme responding. Previous studies showed that, indeed, response styles and traits are often simultaneously involved in certain sub-decisions in empirical data (Meiser et al., 2019; Merhof & Meiser, 2023; von Davier & Khorramdel, 2013). Moreover, multidimensional pseudo-items have the advantage that even if the sub-decisions originate from a unidimensional response process, it is not required to specify in advance which person parameter is driving this decision. Rather, both the dimensionality of sub-decisions and the involved parameters can be explored in the given data.

The IRTree model with multidimensional pseudo-items used in the following is given by

$$p(Y_{vi} = y_{vi}) = \left[ \frac{\exp(x_{1vi}(\theta_v - \beta_{i1}))}{1 + \exp(\theta_v - \beta_{i1})} \right] \times \left[ \frac{\exp(x_{2vi}(\eta_v + \lambda\theta_v - \beta_{i2}))}{1 + \exp(\eta_v + \lambda\theta_v - \beta_{i2})} \right]^{x_{1vi}} \times \left[ \frac{\exp(x_{3vi}(\eta_v - \lambda\theta_v - \beta_{i3}))}{1 + \exp(\eta_v - \lambda\theta_v - \beta_{i3})} \right]^{(1-x_{1vi})} \quad (6)$$

with  $\lambda \geq 0$ .

The model differs from the standard IRTree model with unidimensional pseudo-items only in the parameterization of extreme responding, for which in addition to the ERS  $\eta$ , also the trait  $\theta$  is assumed to influence the respondents' decisions. The weight parameter  $\lambda$  indicates the relative importance of the trait for extreme responding in relation to its importance for the agreement decisions, in which it is weighted by one. The trait is given opposite signs for extreme responding conditional on the agreement judgment to account for the fact that extreme agreement is more likely under both high ERS and high trait levels of respondents, whereas the probability of selecting extreme instead of non-extreme disagreement still increases with higher ERS but decreases with higher trait levels. These differently directed influences of trait and ERS across the extreme pseudo-items facilitate the statistical and meaningful separation of the two person parameters, despite the fact that they do not relate to distinct sub-decisions.

IRTree models with multidimensional pseudo-items can be expected to counteract mimicry effects since the response style parameters are statistically ineffective substitutes for the substantive trait if the trait itself is also included in the respective sub-decisions. As illustrated previously for the IRTree model with unidimensional parameterization (see Figure 2), ERS parameters mimicking the trait are advantageous for explaining the trait-induced variance of extreme responding for one side of the rating scale (e.g., variance among the agreement categories) and disadvantageous for the other side (e.g., variance among the disagreement categories). Only if the responses are asymmetrically distributed over both sides of the rating scale, as is the case for shifted or skewed trait distributions, the model benefits from the redeclaration of the ERS factor. In contrast, since the IRTree model with multidimensional pseudo-items incorporates trait influences for all sub-decisions, the trait parameters can account for the trait-induced variance in extreme responding independently of the response distribution. Multidimensional pseudo-items can thus be assumed to not only maintain the statistical separation of traits and response styles but also enhance the meaningful separation of such parameters in comparison to the unidimensional parameterization.

In the second simulation study, both the IRTree model with multidimensional pseudo-items and the model with fixed covariance were evaluated with regard to mimicry effects and trait recovery.<sup>3</sup> They were compared against the standard IRTree model used in the first simulation study, in which the covariance of trait and ERS was estimated and all pseudo-items were parameterized by unidimensional IRT

models. The same unidimensional data-generating procedure by the PCM as in the first simulation study was applied. As mimicry effects were found to likewise occur for data with shifted and skewed trait distributions (see Table 1), only the shifted conditions were considered here. Furthermore, sample size and questionnaire length were not varied ( $N$  was set to 500, and  $I$  was set to 20), as no relevant differences were observed (see Tables A1 and A2 in the Online Supplementary Materials). 100 replications were conducted for each shifted condition with  $\mu$  set to 0.0, 0.2, 0.5, and 1.0.

## Results

The analysis of the estimated variances and covariances (see Table 3) revealed that the IRTree model with fixed covariance was only suitable to a limited extent in terms of counteracting mimicry effects and the misattribution of trait-induced variance. Although fixing the covariances of ERS and trait naturally prevents mimicry effects in the strict sense, the estimated variances of ERS parameters increased with increasing trait shifts. This overestimation of the ERS variance suggests that the ERS parameters still captured part of the trait-induced variance in extreme responding. We therefore investigated whether, despite the zero-constrained population covariance ( $\widehat{Cov}(\eta, \theta) = 0$ ), the covariance of the estimated trait and ERS levels ( $Cov(\hat{\eta}, \hat{\theta})$ ) nevertheless differed from zero. The covariance of estimated parameters indeed increased with increasing trait shifts, demonstrating that a kind of hidden mimicry effect occurred. As a result, the ERS parameters mimicked the trait, causing them to covary with each other, although the constrained population covariance supposedly specified that there was no relationship between the parameters. As this hidden mimicry effect was smaller compared with the actual mimicry effect that occurred in the standard IRTree model, forcing the population covariance to zero seems to have suppressed at least part of the redeclaration of the ERS parameters (for  $\mu = 1.0$ , the hidden effect was 0.26 and the mimicry effect of the standard IRTree model was 0.69, see Table A1 in the Online Supplementary Materials, condition with  $N = 500$ ,  $I = 20$ ). Nevertheless, the model with fixed covariance did not prevent biases in the parameter estimation to a satisfactory degree, as it still indicated that an ERS influence was present, even though it was not part of the data-generating process.

In contrast, the IRTree model with multidimensional pseudo-items provided estimates of the ERS variance which were very close to zero regardless of the trait distribution. Thus, it successfully detected the unidimensional data-generating process and accurately reflected the absence of response style influences. The covariances of ERS and trait were likewise correctly estimated to be close to zero so that mimicry effects did not occur. The IRTree model with multidimensional pseudo-items therefore consistently prevented a misattribution of the trait-induced variance even for strongly asymmetrical response distributions.

Somewhat unexpectedly, the correlations of the ERS with the trait estimated by the model with multidimensional pseudo-items were on average slightly negative. As

**Table 3.** Estimated Covariances, Correlations, and Variances of ERS  $\eta$  and Trait  $\theta$  by the Modified IRTree Models for Unidimensional Data (Simulation 2).

Analysis	Trait shift	M (SD) across replications				
		$\widehat{\text{Cov}}(\eta, \theta)$	$\text{Cov}(\hat{\eta}, \hat{\theta})$	$\widehat{\text{Cor}}(\eta, \theta)$	$\widehat{\text{Var}}(\theta)$	
IRTree fixed covariance	$\mu = 0.0$	0.00 (0.00)*	0.00 (0.07)	0.00 (0.00)*	0.24 (0.06)	2.09 (0.20)
	$\mu = 0.2$	0.00 (0.00)*	0.05 (0.07)	0.00 (0.00)*	0.24 (0.06)	2.10 (0.23)
	$\mu = 0.5$	0.00 (0.00)*	0.12 (0.07)	0.00 (0.00)*	0.26 (0.07)	2.10 (0.22)
	$\mu = 1.0$	0.00 (0.00)*	0.26 (0.09)	0.00 (0.00)*	0.34 (0.08)	2.13 (0.22)
IRTree multidimensional	$\mu = 0.0$	-0.01 (0.06)	-0.01 (0.05)	-0.05 (0.18)	0.04 (0.01)	2.10 (0.20)
	$\mu = 0.2$	-0.02 (0.06)	-0.02 (0.05)	-0.08 (0.19)	0.04 (0.01)	2.10 (0.23)
	$\mu = 0.5$	-0.02 (0.06)	-0.01 (0.05)	-0.06 (0.20)	0.04 (0.01)	2.10 (0.24)
	$\mu = 1.0$	-0.05 (0.07)	-0.03 (0.06)	-0.14 (0.20)	0.05 (0.01)	2.13 (0.23)

Note.  $N = 500$ ,  $I = 20$ . ERS = extreme response style; SD = standard deviation.

\* Covariance and correlation of ERS and trait are not estimated but fixed to zero.

**Table 4.** Trait Recovery  $\text{Cor}(\theta, \hat{\theta})$  for Unidimensional Data (Simulation 2).

Trait shift	<i>M (SD)</i> across replications			
	IRTree	IRTree fixed coverage	IRTree multidimensional	PCM
$\mu = 0.0$	0.92 (0.01)	0.92 (0.01)	0.95 (0.01)	0.95 (0.01)
$\mu = 0.2$	0.92 (0.01)	0.92 (0.01)	0.95 (0.01)	0.95 (0.01)
$\mu = 0.5$	0.92 (0.01)	0.92 (0.01)	0.95 (0.00)	0.95 (0.00)
$\mu = 1.0$	0.92 (0.01)	0.91 (0.01)	0.94 (0.01)	0.94 (0.01)

Note.  $N = 500$ ,  $I = 20$ . *SD* = standard deviation; PCM = partial credit model.

these estimates largely varied across simulation replications, and given that the individual ERS factors had a very low variance, this correlation is likely an artifact of the estimation and suggests that the parameters adapted to small random variations of the respondents' selection of extreme categories. Furthermore, the fact that the variance of the ERS and its covariance with the trait were consistently estimated to be close to zero for all data sets, the model would hardly mislead to the false interpretation that the small correlations between trait and ERS were substantially meaningful.

Also, the recovery of trait levels (see Table 4) was comparatively better in the IRTree model with multidimensional pseudo-items than in the models with unidimensional ones (with fixed or estimated covariance) and even reached the benchmark recovery by the true data-generating PCM. However, the models with unidimensional pseudo-items performed only slightly worse and still yielded satisfactory recovery. As was shown in the first simulation study, mimicry effects and misspecifications of IRTree models were only of limited relevance for the trait recovery and were more severe in terms of possible incorrect conclusions about the presence and importance of response styles.

Altogether, the second simulation study demonstrated that the model with multidimensional pseudo-items successfully counteracted mimicry effects and further recovered the substantive trait levels very well. These results suggest that a multidimensional parameterization of pseudo-items should be preferred to a unidimensional one if it seems plausible from a theoretical perspective, for instance, if sub-decisions that are assumed to be based on response styles may be additionally influenced by the trait. So far, though, we have provided evidence for the benefits of the multidimensional parameterization only for unidimensional data. However, item responding without any response style influence is (a) hardly found in empirical data and (b) contrary to the primary purpose of using IRTree models, namely to control trait measurements for response style effects. Therefore, a third simulation study was conducted, in which the previously analyzed IRTree models with unidimensional or multidimensional pseudo-items were fitted to data originating from a multidimensional response process with ERS influence.

### Simulation Study 3—Multidimensional Data With Response Style Influence

The third simulation study concerned mimicry effects in multidimensional item response data, for which, in addition to the trait, also a response style was assumed to affect the selection of rating categories. For such data, the mimicry effect is the difference between the true and estimated covariances, which is in contrast to the first two studies where the estimated covariance directly quantified the mimicry effect. Nevertheless, just as for unidimensional data, we assumed that the response style parameters should mimic the trait and capture part of its variance if a misspecified IRTree model overstated the influence of a response style and understated the influence of the trait. For example, an IRTree model could suggest that the extreme sub-decisions were purely ERS-based, although, in fact, the trait additionally affected the response selection. Since a skewed or shifted trait distribution results in an asymmetrical response distribution also for multidimensional data, we expected a statistical advantage of adjusting the meaning of the ERS toward that of the substantive trait. However, unlike in unidimensional data, only part of the estimated response style variance should then reflect trait-based responding, and the other part should reflect actual differences in individual category preferences. In the data example illustrated in Figure 2B, the person-specific ERS estimates should thus represent a compromise between the true preferences for extreme categories and trait-based responding. The balance of this compromise should depend on the response distribution so that a higher asymmetry should cause the estimated ERS levels to more closely reflect the substantive meaning of the trait. Likewise, the estimated covariance of response style and trait should comprise the sum of both the true relationship of the latent personal characteristics and the artificially evoked covariance. Thereby, the direction of the asymmetry can be assumed to determine whether the covariance is overestimated or underestimated, so a mimicry effect may even change the sign of the estimated relationship.

Our hypotheses on mimicry effects in multidimensional data were tested in the simulation study by generating item response data under the IRTree model with multidimensional pseudo-items according to Equation 6, in which the agreement sub-decision is modeled to be solely dependent on the trait, and the extreme sub-decisions are parameterized by both the trait and the ERS.<sup>2</sup> The parameter  $\lambda$ , which indicates the importance of the trait for extreme responding, was set to 0.5 across all generated data sets, as previous studies showed that such is a realistic value for empirical data (Meiser et al., 2019; Merhof & Meiser, 2023). Since the mimicry effect is the covariance of a response style and the trait which deviates from the true relationship of these two parameters, we varied the covariance of ERS and trait as an additional simulation factor and set it to 0.0, 0.2, 0.4, and 0.6. The variances of both ERS and trait were set to 1, so the generated covariances corresponded to the correlations. The trait shift  $\mu$  was varied and set to 0.0 and 1.0. The sample size  $N$  was set to 100, 500, and 2,000; the questionnaire length  $I$  was set to 5, 10, 20, and 40. 100 replications were conducted for each condition of the fully crossed simulation factors. The



**Table 5.** Estimated Covariances and Correlations of ERS  $\eta$  and Trait  $\theta$  for Multidimensional Data With ERS Influence (Simulation 3).

Analysis	Cov( $\eta, \theta$ )	$M$ (SD) across replications			
		$\widehat{\text{Cov}}(\eta, \theta)$		$\widehat{\text{Cor}}(\eta, \theta)$	
		$\mu=0.0$	$\mu=1.0$	$\mu=0.0$	$\mu=1.0$
IRTree	0.0	0.00 (0.16)	0.14 (0.16)	0.00 (0.15)	0.14 (0.16)
	0.2	0.19 (0.16)	0.33 (0.16)	0.19 (0.16)	0.32 (0.14)
	0.4	0.38 (0.16)	0.54 (0.17)	0.38 (0.14)	0.51 (0.13)
	0.6	0.57 (0.17)	0.73 (0.17)	0.58 (0.13)	0.67 (0.10)
IRTree multidimensional	0.0	-0.01 (0.14)	-0.01 (0.14)	-0.01 (0.20)	0.00 (0.19)
	0.2	0.19 (0.14)	0.19 (0.14)	0.20 (0.20)	0.21 (0.18)
	0.4	0.40 (0.14)	0.40 (0.14)	0.42 (0.17)	0.43 (0.18)
	0.6	0.59 (0.15)	0.59 (0.15)	0.62 (0.13)	0.63 (0.13)

Note. Aggregated across sample sizes ( $N = 100, 500, 2,000$ ) and questionnaire lengths ( $l = 5, 10, 20, 40$ ).  
ERS = extreme response style.

analysis models were the standard IRTree model with unidimensional pseudo-items, the IRTree model with multidimensional pseudo-items, and the PCM.

## Results

The analyses clearly demonstrated that mimicry effects also occur if a response style was involved in the data-generating process in addition to the trait. As shown in Table 5, the standard IRTree model with unidimensional pseudo-items yielded inflated estimates of the covariance in case of a trait shift. The true covariance of ERS and trait hardly influenced the mimicry effect, as the overestimation of covariance was consistent across the conditions of generated covariances. However, with an average overestimation of 0.14, the mimicry effect was considerably smaller than the corresponding effect for unidimensional data (0.70, see Table 1). This difference is due to the fact that the ERS parameters only capture the trait-induced variance of extreme responding for unidimensional data but are a compromise of the trait and actual ERS levels of respondents in multidimensional data. As was shown for unidimensional data, sample size and questionnaire length did not influence the mimicry effect (see Table A3 in the Online Supplementary Materials).

The IRTree model with multidimensional pseudo-items again proved to be resistant to the mimicry effect, as it provided unbiased estimates of the covariance across all conditions also for multidimensional data. This finding corroborates our previous suggestion that a multidimensional parameterization of pseudo-items should be preferred to a unidimensional one if unidimensionality is not required from a theoretical point of view. Also in terms of person parameter recovery, a similar pattern to that observed for unidimensional data was found: The differences in recovery of trait and

**Table 6.** Parameter Recovery for Multidimensional Data With ERS Influence (Simulation 3).

Analysis	Cov( $\eta, \theta$ )	<i>M (SD)</i> across replications			
		Cor( $\theta, \hat{\theta}$ )		Cor( $\eta, \hat{\eta}$ )	
		$\mu = 0.0$	$\mu = 1.0$	$\mu = 0.0$	$\mu = 1.0$
IRTree	0.0	0.80 (0.11)	0.79 (0.12)	.78 (.11)	.77 (.11)
	0.2	0.80 (0.11)	0.79 (0.11)	.78 (.11)	.78 (.11)
	0.4	0.80 (0.11)	0.80 (0.10)	.79 (.11)	.79 (.10)
	0.6	0.81 (0.10)	0.82 (0.09)	.80 (.10)	.81 (.09)
IRTree multidimensional	0.0	0.82 (0.11)	0.81 (0.11)	.78 (.13)	.78 (.14)
	0.2	0.82 (0.12)	0.81 (0.11)	.78 (.13)	.79 (.12)
	0.4	0.82 (0.11)	0.82 (0.11)	.79 (.12)	.79 (.12)
	0.6	0.83 (0.10)	0.84 (0.09)	.81 (.11)	.81 (.11)
PCM	0.0	0.81 (0.10)	0.80 (0.10)	—	—
	0.2	0.81 (0.10)	0.81 (0.10)	—	—
	0.4	0.81 (0.10)	0.82 (0.10)	—	—
	0.6	0.81 (0.10)	0.82 (0.09)	—	—

Note. Aggregated across sample sizes ( $N = 100, 500, 2,000$ ) and questionnaire lengths ( $I = 5, 10, 20, 40$ ). ERS = extreme response style; SD = standard deviation; PCM = partial credit model.

ERS levels between the models were small, with a slight advantage of the true data-generating IRTree model with multidimensional pseudo-items (see Table 6). Only in conditions with few data points ( $N = 100$  and  $I = 5$ ), the recovery by the IRTree model with multidimensional pseudo-items was slightly worse compared with the other models, which is probably due to the greater complexity of this model (see Tables A4 and A5 in the Online Supplementary Materials for the recovery of person parameters split by  $N$  and  $I$ ).

## Application

To demonstrate the impact of mimicry effects on the validity of conclusions drawn from empirical data, two scales of the background questionnaire of the PISA 2018 study were analyzed by IRTree modeling. We used the item responses of  $N = 4,411$  participants to the 2 scales “reading self-evaluation” comprising 6 items and “reading enjoyment” comprising five items on a 4-point rating scale.<sup>4</sup> The subset of the data considered here is described in more detail by Henninger and Meiser (2023). The standard IRTree model with unidimensional pseudo-items as well as the IRTree model with multidimensional pseudo-items were fitted to the data. As the multidimensional model was shown to produce unbiased estimates in the simulation studies, it was considered the benchmark model with which the standard IRTree model was compared in order to quantify mimicry effects.

First, both scales were analyzed separately for illustration purposes. The results are summarized in Table 7 and suggest mimicry effects in the standard IRTree model

**Table 7.** Estimated Covariances, Correlations, and Variances of ERS  $\eta$  and Trait  $\theta$  for the Empirical PISA Data.

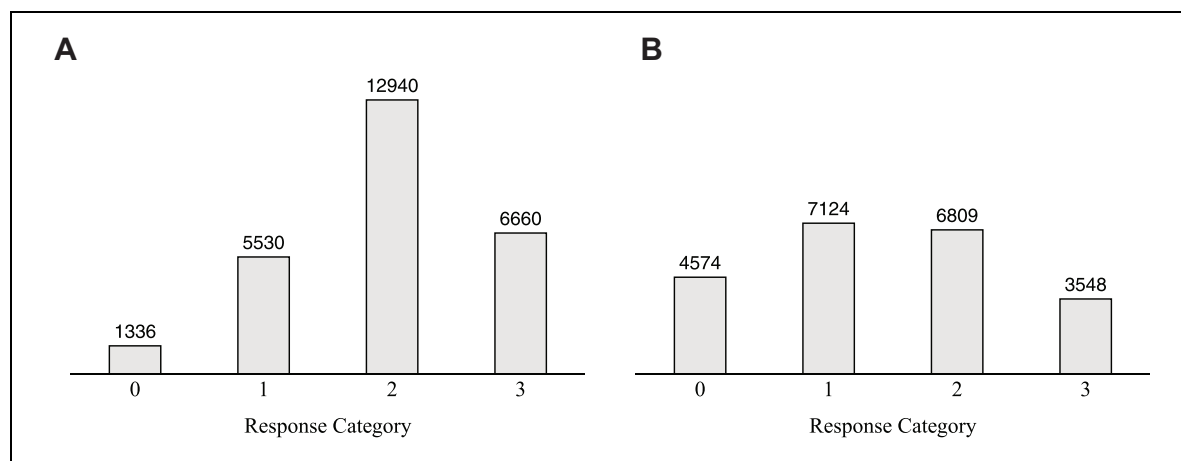
PISA scale	Analysis	$\widehat{\text{Cov}}(\eta, \theta)$	$\widehat{\text{Cor}}(\eta, \theta)$	$\widehat{\text{Var}}(\eta)$	$\widehat{\text{Var}}(\theta)$
Reading	IRTree	1.92	0.41	5.74	3.87
self-evaluation	IRTree multidimensional	0.72	0.19	4.02	3.63
Reading	IRTree	0.12	0.02	3.38	7.10
enjoyment	IRTree multidimensional	0.49	0.13	2.16	7.05

Note. ERS = extreme response style; PISA = Program for International Student Assessment.

for both scales: The estimated covariances, correlations, and ERS variances largely differed between the two models, indicating a biased estimation by the IRTree model with unidimensional pseudo-items. For “reading self-evaluation,” the correlation of ERS and trait under the standard IRTree model was of substantial size, which was strongly reduced when the model with multidimensional pseudo-items was applied. For “reading enjoyment,” a mimicry effect was likewise apparent, though in the opposite direction: The model with unidimensional pseudo-items indicated that trait and ERS levels were unrelated, when in fact the model with multidimensional pseudo-items showed that they were positively correlated. In both cases, the larger ERS variances in the standard IRTree model supported the presence of mimicry effects.

Thereby, the distributions of the observed item responses (see Figure 3) further demonstrated that even a slight asymmetry toward one side of the rating scale can distort the interpretation of the results derived from the IRTree model with unidimensional pseudo-items: Whereas the item responses of the “reading self-evaluation” scale reveal a noticeable asymmetry toward the agreement side of the scale, the distribution of “reading enjoyment” appears to be rather symmetrical. Since erroneous parameter estimates nevertheless occurred for both scales, this application example highlights that visual inspections of observed item responses are not necessarily indicative of mimicry effects, and should not be used as the sole diagnostic criterion for choosing the IRTree model applied to the data.

In an additional analysis, both scales were modeled simultaneously, which is generally preferable to the separate analysis of multiple scales. Since a joint model, which defines a response style as a category preference across several unrelated constructs, facilitates separating response style and trait factors more accurately (e.g., Bolt & Newton, 2011), this approach should also reduce mimicry effects. For the PISA data, the mimicry effects were indeed reduced in the joint model: The difference in the estimated correlation between trait and ERS by the IRTree model with unidimensional versus multidimensional pseudo-items decreased from .22 to .19 for “reading self-evaluation” and from .11 to .08. for “reading enjoyment.” The fact that mimicry effects were still present and of substantial size, however, is likely due to the high correlation of the two content traits of .48. Therefore, also when jointly



**Figure 3.** Response Distributions and Absolute Category Frequencies Across the Items of the Two Scales in the Empirical PISA Data. (A) Reading Self-Evaluation. (B) Reading Enjoyment.

Note. PISA = Program for International Student Assessment.

analyzing multiple scales, IRTree models with multidimensional pseudo-items should be considered.

## Discussion

This article investigated the separation of substantive traits and response styles in IRTree models and addressed the threat of mimicry effects, a methodological artifact where response styles mimic the trait and capture trait-induced variance in item responding. As the response style factor functions as a substitute for the trait in such instances, the meaning of the estimated response styles does not correspond to the meaning that was assigned to the parameters when defining the model. Mimicry effects are manifested in a biased estimation of the covariance between response style and trait, with the bias being stronger the more the meanings of the two factors overlap. The covariance can be overestimated as well as underestimated, both of which can lead to severely misleading conclusions about the relationship between personal characteristics. For example, the IRTree model estimates may suggest that high levels of the trait of interest are associated with preferences for specific categories, although there is no or even an opposite relationship between these parameters. In addition to the biased estimation of the covariance, mimicry effects were found to be accompanied by inflated estimates of response style variances, meaning that the impact of a response style on the response selection is overestimated. In extreme cases, IRTree models might even misjudge the dimensionality of the data-generating process and indicate an influence of response styles where respondents actually provided purely trait-based responses. It could thus be concluded that some respondents did not work on the questionnaire with full effort but relied heavily on their response styles, although they engaged in an optimal and desired way of response selection. Particularly when dealing with high-stakes data such as assessments in job

interviews, the false assumption that some applicants have made little effort to complete the task comes with potentially negative implications for such individuals and jeopardizes fairness. Consequently, it is important for both research and practice to be aware of possible methodological artifacts in IRTree models and to question the assigned meaning of estimated parameters rather than to interpret them as substantially meaningful without further consideration.

### *Conditions and Implications of Mimicry Effects*

An important research question is, therefore, under which conditions IRTree models pose the risk of artificial estimates. Our investigations suggested that only those IRTree models evoke mimicry effects, which are misspecified in a way that they overstate the influence of response styles and understate the influence of the trait in some pseudo-items. For example, this is the case if the item responses of a given data set originated from a unidimensional trait-based process, though an IRTree model with trait and ERS influences is applied. In such cases, the ERS factor can be used as a substitute for the trait and explain the trait-induced variance in extreme responding; in other words, a mimicry effect arises. In addition, the simulation studies corroborated our hypothesis that mimicry effects are largely dependent on the distribution of ordinal item responses across the rating scale. If they are symmetrically distributed with similar frequencies of agreement and disagreement categories, unbiased estimates of the variance-covariance matrix are provided. As such symmetrical response patterns yield no statistical advantage of a redeclaration of the response style parameter as a substitute for the trait, mimicry effects do not occur. In contrast, the more asymmetrically the responses are distributed across the scale, the better the variance in extreme responding can be explained if the response style parameters mimic the trait and capture trait-induced variance. As a result, mimicry effects occur and the meaningful separation of trait and response style parameters is compromised.

In the simulation studies, we operationalized the asymmetry of item responses by generating distributions of the trait levels which deviated from those of the item locations, through specifying either shifted mean structures or skewed distributions. Both led to considerable mimicry effects and an overestimation of the impact of response styles, even for small deviations between the distributions. This finding is highly relevant, as it is certainly not uncommon to apply a questionnaire to a group of individuals for whom it can be assumed that their traits are at least slightly differently distributed from that of the items. The empirical application to PISA data supported the results of the simulations, as indeed, even a small asymmetry of the observed responses was found to result in a mimicry effect.

Besides the threat of biased interpretations when analyzing the data of a single group of respondents, mimicry effects may likewise distort comparisons between multiple groups if such differ in their trait distributions. An example is cross-national assessments, for which one would certainly expect group differences in the distributions of the measured constructs, which would cause also the size of possible mimicry

effects to vary. Though the comparisons of the trait of interest would not be impaired by mimicry effects, one may conclude that the groups differed in the extent of using response style (e.g., supposedly caused by different cultural backgrounds and socializations). IRTree models should thus be used with caution when shifted or skewed trait distributions may be present in the data, which is likely the case for many applications across all fields of psychology in which self-reported data are analyzed (e.g., clinical, personality, or work psychology).

However, this article also showed that not all types of IRTree models were prone to mimicry effects. The concerns and criticisms outlined above referred to the commonly used IRTree models in which all pseudo-items are parameterized by unidimensional models. Such models are based on the assumption that each sub-decision is affected by only one personal characteristic, which can be the trait or a response style. A different assumption is underlying IRTree models with multidimensional pseudo-items, in which the sub-decisions can be assigned several person parameters (e.g., the trait plus a response style). The simulation studies demonstrated that if the trait is additionally included in a pseudo-item, in which a response style would mimic the trait in the standard IRTree model with unidimensional parameterization, the trait itself accounts for the trait-induced variance, and the mimicry effect is prevented. The ability of such IRTree models to counteract mimicry effects was apparent in all simulation conditions of generated trait distribution, that is, was independent of the symmetry or asymmetry of the response distribution.

Furthermore, the advantage of a multidimensional parameterization of pseudo-items was not only evident for unidimensional, trait-based data-generating processes but also for more realistic multidimensional ones. We generated data under a two-dimensional IRTree model, in which the extreme pseudo-items were influenced by the ERS and the trait. Regardless of the true covariance of response style and trait, the IRTree model with multidimensional pseudo-items provided unbiased estimates and accurately reflected their true relationship. In contrast, the standard IRTree model with unidimensional pseudo-items led to mimicry effects whenever the response distribution was asymmetrical, although the size of such mimicry effects for multidimensional data was smaller compared with the effects in unidimensional data. This comparatively smaller mimicry effect indicates that the response style parameters are used to capture variance of both trait-based and response style-based responding for multidimensional data, and therefore, have less overlap with the trait compared with unidimensional data. Even though the potential for misinterpretations was consequently less severe under the more realistic multidimensional data, a disadvantage of unidimensional pseudo-items compared with multidimensional ones was still evident.

Despite improved psychometric properties of the models with multidimensional pseudo-items, the simulation studies also showed that the trait recovery was hardly affected by mimicry effects and biased response style estimates. Accordingly, the main purpose of response style modeling in empirical research and practice, namely, to obtain unbiased trait measurements, was successfully realized by both unidimensional and multidimensional parameterizations. Nevertheless, applying a model that

yields biased estimates under certain circumstances of misspecification, even if such parameters are not of interest, should generally be avoided, as such a model is unable to provide information on the true data-generating process.

### *Recommendations for the Specification of IRTree Models*

Therefore, this article provides some suggestions on how to specify IRTree models and how to adapt them to the given research question and data: First, knowledge about the construct to be measured and about the questionnaire that is applied helps to anticipate whether the distributions of items and traits are likely to match or deviate from each other. Such theoretical considerations give an indication of whether using a standard IRTree model with unidimensional pseudo-items carries a risk of mimicry effects even before the data are collected. After data collection, it should be examined whether the empirical distribution of the item responses is symmetrical or asymmetrical. However, the application example demonstrated that even a slight asymmetry of responses, which can be easily overlooked or considered negligible, can lead to mimicry effects and change the interpretation of results. Unexpectedly high correlations of traits and response styles could thus be regarded as a warning sign for a possible mimicry effect. Nonetheless, mimicry effects can likewise result in an artificial reduction of an estimated relationship, which is probably a less obvious warning sign. We therefore recommend that IRTree models with unidimensional pseudo-items should only be applied if the trait distribution matches that of the items well, or if the response style estimates and the relationships between person parameters are not of interest for answering the research question. Of course, there may be certain hypotheses to be tested that require the specification of unidimensional processes, or only purely unidimensional sub-decisions are reasonable from a theoretical point of view. In such cases, it could be advisable to define an IRTree model across several questionnaire scales, though the benefits may be limited if the traits are correlated, as was evident in the application example. Therefore, further investigations may be needed to clarify how and to what extent the occurrence of mimicry effects can be reduced by simultaneously modeling several traits.

As a result, our analyses indicate that a multidimensional parameterization of pseudo-items should be generally preferred to a unidimensional one whenever possible. The advantage of multidimensional pseudo-items is all the more apparent since a possible overparameterization (e.g., using a two-dimensional parameterization for unidimensional pseudo-items) has no negative effect on the parameter estimation, as a non-existent influence of one of the person parameters is successfully detected by IRTree models. Moreover, the sub-decisions may actually be the result of a multidimensional response process, in which case only multidimensional pseudo-items can correctly reflect the true data-generating process. We thus believe that the prevention of mimicry effects and the greater flexibility of multidimensional parameterizations of pseudo-items outweigh the slightly higher modeling complexity in comparison to unidimensional pseudo-items. Furthermore, multidimensional pseudo-items can be

readily implemented in standard software with little additional effort (like in the R package *mirt*; see the Online Supplementary Materials for *mirt* code of various IRTree models).

## Outlook

One limitation of this work is that we only considered one response style, the ERS, which is one of the most studied response styles in the literature. However, mimicry effects are probably also relevant for modeling other types of response styles such as the MRS. A classical IRTree model including MRS-based judgments defines three sub-decisions for responding to 6-point items, which are the decisions of agreement, moderate responding, and extreme responding (e.g., Böckenholt, 2017; Meiser et al., 2019): Respondents are assumed to first decide on whether they agree or disagree with the item and subsequently make an MRS-based sub-decision for midscale versus non-midscale responding conditional on agreement. In case they chose the non-midscale option, they decide on the extremity of their response based on their ERS. Just as derived before for the ERS, the meaning of the MRS is separated from that of the trait by defining a unique influence of the MRS across two pseudo-items (conditional on agreement and disagreement). Shifted or skewed trait distributions and asymmetrical response distributions should therefore most likely impair the separation of trait and MRS parameters and lead to mimicry effects also for the MRS. Still, the midscale pseudo-items can likewise be parameterized by multidimensional IRT models including an additional trait influence, which should counteract mimicry effects as successfully as shown here for the ERS. Although mimicry effects are thus likely to generalize to other response styles, it nevertheless remains to be clarified how they affect the parameter estimation of IRTree models when several response styles (e.g., ERS and MRS) are jointly modeled.

Furthermore, this article investigated mimicry effects only in IRTree models with a symmetrical tree structure (also called nested IRTree models), in which the same sequence of response processes is assumed to underlie the selection of corresponding categories on both sides of the rating scale. However, IRTree models can also be defined to have an asymmetrical structure (for an overview of different kinds of IRTree models, see Jeon & De Boeck, 2016). An example of such an asymmetrical IRTree model is the commonly used decomposition of 5-point rating items, in which one sub-decision represents the MRS-based choice to select either the neutral middle response category or one of the other categories. Conditional on the selection of a clear-cut category, two sub-decisions of trait-based agreement and ERS-based extreme responding are specified (Böckenholt, 2012; also see Khorramdel & von Davier, 2014; Plieninger, 2020; Zettler et al., 2016). In contrast to the models used in this article, the MRS is separated from the trait by means of only one pseudo-item. Such a model structure can be expected to likewise lead to mimicry effects for asymmetrical response distributions, though this should be systematically investigated and quantified and future work.



Various other modeling choices have been made in this article, the findings of which can be straightforwardly generalized to other choices: First, all IRTree models considered here were parameterized by the Rasch model, though other IRT models such as the 2PL model should naturally lead to similar results regarding mimicry effects. Furthermore, challenges in the separation of person parameters can be expected to not only occur for shifted or skewed trait distributions but also if traits and items mismatch otherwise. An example is bimodal trait distributions, which could result from an unknown mixture of two populations. Finally, mimicry effects can even be generalized beyond the IRTree model class to multidimensional ordinal IRT models such as the multidimensional nominal response model or the multidimensional PCM (e.g., Bolt et al., 2014; Falk & Cai, 2016 for an overview, see Henninger & Meiser, 2020). Just as for IRTree models, the meaningful separation of traits and response styles in such models is facilitated through uniquely directed influences of the person parameters, which can be assumed to be impaired if the distributions of the trait levels and item locations deviate from each other.

Overall, this article presented compelling evidence for the risk of mimicry effects in commonly used IRTree models. To address these concerns, we made suggestions on how to detect the lack of meaningful separation of traits and response styles and showed that IRTree models with multidimensional pseudo-items effectively counteract such mimicry effects. Our findings highlight the importance of being aware of potential methodological artifacts when modeling item response data and underline that further research is needed to ensure the validity of conclusions drawn from such data.

### **Acknowledgment**

The authors thank Fabiola Reiber for valuable discussions and helpful suggestions regarding the simulation studies.

### **Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group Statistical Modeling in Psychology (SMiP).

### **ORCID iD**

Viola Merhof  <https://orcid.org/0000-0002-1328-0000>

## Data Availability Statement

The results of individual replications of the simulation studies can be found on OSF: <https://osf.io/9raud/> The data used for reanalyses in the empirical application are available here: <https://www.oecd.org/pisa/data/2018database/>

## Supplemental Material

Supplemental material for this article is available online.

## Notes

1. We also compared the recovery of item difficulty parameters, although this is of limited relevance for most practical applications. The results of the item difficulty recovery for all simulation studies can be found in the Online Supplementary Materials.
2. The R code for generating data sets can be found on OSF: <https://osf.io/9raud/>.
3. The *mirt* code for such models can be found in the Online Supplementary Materials.
4. The data are openly available here: <https://www.oecd.org/pisa/data/2018database/>. The background questionnaire with all items can be found here: [https://nces.ed.gov/surveys/pisa/pisa2018/questionnaires/Student\\_Q\\_Booklet\\_English.html](https://nces.ed.gov/surveys/pisa/pisa2018/questionnaires/Student_Q_Booklet_English.html).

## References

- Azzalini, A., & Capitanio, A. (2014). *The skew-normal and related families*. Cambridge University Press.
- Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Beck, A. T., Steer, R. A., & Brown, G. (1996). *Beck's Depression Inventory—II*. <https://naviauxlab.ucsd.edu/wp-content/uploads/2020/09/BDI21.pdf>
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17(4), 665–678. <https://doi.org/10.1037/a0028111>
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22(1), 69–83. <https://doi.org/10.1037/met0000106>
- Böckenholt, U. (2019). Assessing item-feature effects with item response tree models. *British Journal of Mathematical and Statistical Psychology*, 72(3), 486–500. <https://doi.org/10.1111/bmsp.12163>
- Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, 70(1), 159–181. <https://doi.org/10.1111/bmsp.12086>
- Bolt, D. M., Lu, Y., & Kim, J.-S. (2014). Measurement and control of response styles using anchoring vignettes: A model-based approach. *Psychological Methods*, 19(4), 528–541. <https://doi.org/10.1037/met0000016>
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, 71(5), 814–833. <https://doi.org/10.1177/0013164410388411>

- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. <https://doi.org/10.18637/jss.v048.i06>
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48(1), 1–28. <https://doi.org/10.18637/jss.v048.c01>
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 21(3), 328–347. <https://doi.org/10.1037/met0000059>
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84–96. <https://doi.org/10.1016/j.jrp.2005.08.007>
- Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, 25(5), 560–576. <https://doi.org/10.1037/met0000249>
- Henninger, M., & Meiser, T. (2023). Quality control: Response style modeling. In R. J. Tierney, F. Rizvi & K. Ercikan (Eds.), *International encyclopedia of education* (pp. 331–340). Elsevier. <https://doi.org/10.1016/b978-0-12-818630-5.10041-7>
- Henninger, M., & Plieninger, H. (2020). Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement. *Assessment*, 28(5), 1301–1319. <https://doi.org/10.1177/1073191119900003>
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48(3), 1070–1085. <https://doi.org/10.3758/s13428015-0631-y>
- Khorramdel, L., & von Davier, M. (2014). Measuring response styles across the Big Five: A multi-scale extension of an approach using multinomial processing trees. *Multivariate Behavioral Research*, 49(2), 161–177. <https://doi.org/10.1080/00273171.2013.866536>
- Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, 81(1), 131–154. <https://doi.org/10.1177/0013164420913915>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- 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 and Statistical Psychology*, 72(3), 501–516. <https://doi.org/10.1111/bmsp.12158>
- Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, 88(4), 1354–1380. <https://doi.org/10.1007/s11336-023-09901-0>
- Moshagen, M., Hilbig, B. E., & Zettler, I. (2018). The dark core of personality. *Psychological Review*, 125(5), 656–688. <https://doi.org/10.1037/rev0000111>
- Plieninger, H. (2020). Developing and applying IR-Tree models: Guidelines, caveats, and an extension to multiple groups. *Organizational Research Methods*, 24(3), 654–670. <https://doi.org/10.1177/1094428120911096>
- Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, 74(5), 875–899. <https://doi.org/10.1177/0013164413514998>

- Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert Scales. *Behavior Research Methods*, *50*(6), 2325–2344. <https://doi.org/10.3758/s13428-017-0997-0>
- 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*(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 463–487). Springer. <https://doi.org/10.1007/978-1-4614-9348-830>
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*(1), 96–110. <https://doi.org/10.1037/a0018721>
- 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*(2), 178–189. <https://doi.org/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*(3), 279–291. <https://doi.org/10.1177/1073191115583714>
- Zettler, I., Lang, J. W. B., Hülshager, 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*(4), 461–472. <https://doi.org/10.1111/jopy.12172>
- Zhang, Y., & Wang, Y. (2020). Validity of three IRT models for measuring and controlling extreme and midpoint response styles. *Frontiers in Psychology*, *11*, Article 271. <https://doi.org/10.3389/fpsyg.2020.00271>



## DYNAMIC RESPONSE STRATEGIES: ACCOUNTING FOR RESPONSE PROCESS HETEROGENEITY IN IRTREE DECISION NODES

VIOLA MERHOF<sup>id</sup> AND THORSTEN MEISER<sup>id</sup>

UNIVERSITY OF MANNHEIM

It is essential to control self-reported trait measurements for response style effects to ensure a valid interpretation of estimates. Traditional psychometric models facilitating such control consider item responses as the result of two kinds of response processes—based on the substantive trait, or based on response styles—and they assume that both of these processes have a constant influence across the items of a questionnaire. However, this homogeneity over items is not always given, for instance, if the respondents' motivation declines throughout the questionnaire so that heuristic responding driven by response styles may gradually take over from cognitively effortful trait-based responding. The present study proposes two dynamic IRTree models, which account for systematic continuous changes and additional random fluctuations of response strategies, by defining item position-dependent trait and response style effects. Simulation analyses demonstrate that the proposed models accurately capture dynamic trajectories of response processes, as well as reliably detect the absence of dynamics, that is, identify constant response strategies. The continuous version of the dynamic model formalizes the underlying response strategies in a parsimonious way and is highly suitable as a cognitive model for investigating response strategy changes over items. The extended model with random fluctuations of strategies can adapt more closely to the item-specific effects of different response processes and thus is a well-fitting model with high flexibility. By using an empirical data set, the benefits of the proposed dynamic approaches over traditional IRTree models are illustrated under realistic conditions.

**Key words:** response styles, item response theory, multidimensional IRTree, item position effects.

Likert-type rating scales are widely used to assess personality, attitudes, or beliefs via self-reports. However, the validity of such trait measurements is threatened by response styles (RS)—tendencies to systematically respond to items on some basis other than what the items were designed to measure (Paulhus, 1991). RS comprise, for instance, preferences for the extreme categories (extreme RS; ERS) or the middle category of the scale (midpoint RS; MRS), irrespective of item content (for an overview, see Van Vaerenbergh & Thomas, 2013). Since RS can systematically bias estimates of substantive traits, resulting in inflated or underestimated individual scores, group means, and correlations of constructs, RS must be controlled for to ensure a valid interpretation of results (Alwin, 2007; Baumgartner & Steenkamp, 2001).

Various item response theory (IRT) approaches have been proposed that facilitate such control of RS effects under conditions in which the underlying response processes are homogeneous across persons and over items, thus assuming a stable response strategy over the course of a questionnaire (e.g., Böckenholt 2017; Bolt & Newton, 2011; Henninger & Meiser, 2020; Plieninger & Meiser, 2014; Wetzels & Carstensen, 2017). Extensions of such models can further account for some kind of heterogeneity of response processes over discrete conditions, either with a focus on latent classes of respondents (e.g., Kim & Bolt, 2021; Tilmstra et al., 2018; von Davier &

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s11336-023-09901-0>.

The data generated in the simulation studies are available in the Open Science Framework repository, <https://osf.io/kc8ve/>. The data sets used for reanalysis in the empirical application are made available by the original author in the Open Science Framework repository, <https://osf.io/tbmh5/>.

Correspondence should be made to Viola Merhof, Department of Psychology, University of Mannheim, L 13 15, 68161 Mannheim, Germany. Email: [merhof@uni-mannheim.de](mailto:merhof@uni-mannheim.de)

Yamamoto, 2007) or with a focus on within-person changes across measurement occasions (e.g., Ames & Leventhal, 2021; Colombi et al., 2021; Weijters et al., 2010). Other approaches include unsystematic item-by-item fluctuations of response strategies within persons (e.g., Plieninger & Heck, 2018; Tijnstra & Bolsinova, in press; Ulitzsch et al., 2022). Neglected so far are *systematic* changes of response strategies *within* a measurement occasion and the associated heterogeneity regarding the manifestations of substantive traits and RS on the level of single items. We aim to close this research gap by modeling dynamic, item position-dependent influences of trait-based and RS-based response processes.

### 1. Dynamic Trait and Response Style Effects

Whenever respondents are asked to provide subjective self-reports by responding to a Likert-type item, they are faced with the challenge of choosing one of several available categories. A prominent theory describing two competing response strategies to bring about such decisions is the conceptualization of Optimizing and Satisficing by Krosnick (1991). According to this framework, the currently applied response strategy depends on the respondents' cognitive effort expended on item responses. On the one hand, giving accurate trait-based responses requires a substantial amount of effort, as four cognitive stages must be proceeded through. These are (1) comprehension of the item, (2) memory search for relevant information, (3) integration of pieces of information into a judgment, and (4) selecting a response category (Tourangeau et al., 2000). Responses derived from such processing are considered optimal, as they are accurate and strong indicators of the true trait levels. On the other hand, if respondents process some or all of the stages heuristically, item responses require less cognitive effort; they are not optimal, but still satisfactory from the respondents' perspective (called satisficing responses). Unlike optimized, solely trait-based responses, such a satisficing response strategy is susceptible to the influence of RS (Aichholzer, 2013; Podsakoff et al., 2012). For instance, respondents may reach the global decision to agree or disagree with an item based on their trait level, but then do not consider the fine nuances between different options that indicate (dis)agreement. In such cases, individual category preferences determine the selection, so that extreme categories are chosen more often by respondents with high ERS levels, whereas midpoint responses are fostered by high levels of MRS. Metaphorically speaking, the decision vacuum left by a superficial instead of thorough trait-based selection is filled by RS-based processes. We, therefore, define *response strategy*, in the narrower sense, as a certain composition of trait-based response processes on the one side and heuristic processes related to one or several RS on the other side.

Whether predominantly trait-based or rather RS-based response strategies are used depends on the cognitive effort that respondents are able and willing to expend on the task, which in turn can be attributed to several properties of items and respondents (for an overview, see Podsakoff et al., 2012): For instance, low respondents' abilities (e.g., low cognitive/verbal ability or education) and high task difficulty (e.g., a complex, abstract, or ambiguous item) can prevent the use of the optimizing, trait-based response strategy (Baumgartner & Steenkamp, 2001; Knowles & Condon, 1999; Krosnick, 1999; Messick, 1991; Podsakoff et al., 2003). Further, various properties of the measurement method (e.g., scale formats or contexts of data collection) were found to affect the degree of response style-related responding (DeCastellarnau, 2018; Van Vaerenbergh & Thomas, 2013). But even if a questionnaire is constructed and applied in a way that respondents are *able* to give optimized responses, insufficient motivation and fatigue can strengthen the RS influence and reduce the quality of responses (Galesic, 2006; Galesic & Bosnjak, 2009; Herzog & Bachman, 1981; Kahn & Cannell, 1957).

Whereas properties of the questionnaire and the response format can be considered fairly homogeneous and unsystematically varying across items, due to careful item construction and

randomization, the respondents' motivation for pursuing the high cognitive effort for optimizing responses may systematically change over time. In line with this, Krosnick (1991) states that "respondents are likely to satisfy whatever desires motivate them to participate just a short way into an interview, and they are likely to become increasingly fatigued, disinterested, impatient, and distracted as the interview progresses" (p. 214). Indeed, item responses are perceived as increasingly burdensome throughout a questionnaire (Galesic, 2006). In addition, long surveys and items presented in later parts of questionnaires reveal a lower data quality with more frequent omissions, dropouts, and response patterns indicating careless responding (Bowling et al., 2021a; Deutskens et al., 2004; Galesic & Bosnjak, 2009; Liu & Wronski, 2018; Marcus et al., 2007). Response times likewise indicate declining test-taking effort and a shift towards heuristic processing: They were found to be shorter for items presented toward the end of a questionnaire (Galesic & Bosnjak, 2009; Yan & Tourangeau, 2008), and such fast responses are associated with less motivation (Bowling et al., 2021b; Callegaro et al., 2009), satisficing responses in general (Andersen & Mayerl, 2017; Zhang & Conrad, 2014), and even more notably, responses that match the person-specific RS (Henninger & Plieninger, 2020). Thus, conditional on a substantial length of a questionnaire, respondents are likely to decrease their investment of cognitive capacity, and rather fall back to fast, heuristic processing. Such dynamic shifts in the response strategy result in a decreasing influence of the substantive trait, while the influence of RS increases over item position.

## 2. Modeling Heterogeneity of Response Processes

The hypothesized dynamic influences of trait-based and RS-based processes reflect a within-person heterogeneity across the items of a questionnaire. There is a wide range of psychometric approaches accounting for heterogeneity in response processes with regard to RS, whereby the distinction between trait-based and RS-based processes has mainly been considered on the between-person level. For instance, mixture Rasch models (e.g., Austin et al., 2006; Gollwitzer et al., 2005; Meiser & Machunsky, 2008), mixture IRTree models (e.g., Khorramdel et al., 2019, Kim & Bolt, 2021), and a general mixture IRT model (Tijmstra et al., 2018) were proposed, which all can be used to identify latent classes of respondents who provide item responses based on different processes, such as responses influenced by response styles or not (i.e., solely trait-based responses). A limitation of such models is that the response process heterogeneity is strictly related to between-person effects so that possible class switches cannot be detected.

Other approaches allow to investigate the within-person stability of RS and to detect changes of respondents' RS levels across discrete measurement occasions, like latent-state-trait models (Weijters et al., 2010; Wetzel et al., 2016), or longitudinal IRTree models (Ames & Leventhal, 2021). A stronger focus on heterogeneous response processes rather than on changes of RS levels per se is provided by hidden Markov models, in which respondents are assumed to hold one of several discrete latent states associated with a particular type of response process, and in which the assignment of respondents to states can change dynamically over measurement occasions (see Kelava & Brandt, 2019). For instance, Colombi et al. (2021) analyzed longitudinal item response data and defined two states, responding with or without the influence of RS, with part of the respondents modeled to freely switch between the two states. Similarly, Ulitzsch et al. (2022) proposed a response time-based mixture model, in which each response of a person is assumed to be stemming from either a careless or an attentive status. Furthermore, heterogeneity at the level of individual items was incorporated in some multi-process models, in which certain decisions during the selection of response categories are assumed to be based on one of several cognitively distinct processes (Plieninger & Heck, 2018; Thissen-Roe & Thissen, 2013; Tijmstra & Bolsinova, in press). For example, in the model by Plieninger and Heck (2018), affirmative



responses can be either an expression of acquiescence RS or of trait-based agreement with the item content, though without accounting for systematically changing strategies.

Taken together, the past research linking RS modeling with heterogeneity of response processes within and between persons has mainly focused on: (1) discrete instead of continuous subpopulations or response states, (2) RS as an attribute that respondents may or may not have, instead of treating them as one of several processes that respondents can use to varying degrees, and (3) heterogeneity between measurement occasions or groups of items, instead of changes on the level of individual items.

In contrast, higher interest in continuous changes of response strategies within a measurement situation exists in item response modeling outside the RS literature. In the research field of performance decline, which describes a decreasing probability of correct responses for achievement items at the end of a test (for an overview, see List et al., 2017), the gradual process change model by Wollack and Cohen (2004) and Goegebeur et al. (2008) is a prominent model for generating and analyzing smooth changes in response strategies (e.g., Huang, 2020; Jin & Wang, 2014; Shao et al., 2016; Suh et al., 2012). In their approach, the response process of random guessing gradually takes over from trait-based problem-solving, and linear as well as curvilinear trajectories can be captured. In a later section of this article, we will account for shifts from effortful to more and more heuristic responses in a similar way, but instead of modeling random guessing for binary performance items, we model ordinal self-ratings and define heuristic responses as strongly influenced by RS.

Thereby, we aim to tackle the previous limitation of RS modeling, being that systematic within-person heterogeneity over the items of a questionnaire was not accounted for. Ignoring shifts in response processes is not only a potential problem for measuring and interpreting person and item parameters, as the dynamic changes themselves can also be the focus of interest: Measures of changes in trait and RS involvement can be used as a diagnostic tool to evaluate questionnaires with regard to the associated burden and required effort, and to compare, for example, subgroups of respondents (e.g., different age groups), subsets of items (e.g., positively and negatively worded items), or modes of data collection (e.g., online vs. lab). Furthermore, a formal model that describes dynamic response strategies can help to understand the interplay of cognitive processes that underlie item responses and to shed light on how respondents arrive at their judgments and decisions. Therefore, we not merely aim to control trait estimates for RS effects but also to provide a cognitive model accounting for dynamic response processes across items.

The remainder of this article is structured as follows: Firstly, traditional IRTree models are introduced. Then, a new dynamic IRTree model for continuously shifting influences of trait-based and RS-based processes is derived and evaluated by a first simulation study. Subsequently, a more flexible, non-continuous version of this model is introduced and likewise tested by a second simulation study. An empirical example is used to demonstrate the benefits of the dynamic approach under realistic conditions. Lastly, the results are interpreted and discussed in light of both basic and applied fields of research.

### 3. IRTree Model Parameterizations of Traits and Response Styles

Multi-process IRTree models (Böckenholt, 2012; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016) decompose response alternatives of rating scales into a sequence of binary pseudo-items, which represent the decisions assumed to be taken by the respondents during item responses. By assigning different latent traits to the pseudo-items, their effects on response selection can be separated. Typically, one pseudo-item represents the decision to agree vs. disagree with the item content, which is supposed to be made based on the substantive trait, whereas all further pseudo-items relate to RS-based responding, like the judgment to give



extreme vs. non-extreme responses guided by ERS (e.g., Böckenholt, 2017; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Zettler et al., 2016).

### 3.1. Unidimensional Node Parameterization

In the following sections, we refer to items on a four-point Likert scale, and we decompose the ordinal item responses into decision nodes of broad agreement and fine-grained extreme responding based on the tree structure depicted in the upper part of Fig. 1. The probability of the ordinal response  $X_{pi} \in \{1, \dots, 4\}$ , representing the categories “strongly disagree”, “disagree”, “agree”, and “strongly agree” of person  $p = 1, \dots, N$  to item  $i = 1, \dots, I$ , is the product of the probabilities of responses to the two pseudo-items  $Y_{hpi} \in \{0, 1\}$  of agreement ( $h = 1$ ) and extreme responding ( $h = 2$ ). This model structure serves as an exemplary illustration for our new approach; dynamic response strategies can be easily adapted to differently structured trees and response formats with more or fewer ordinal categories (see Sect. 7 for an extension to five-point Likert-type items).

In the frequently applied Rasch IRTree, the two pseudo-items are each parameterized by a dichotomous Rasch model, with the agreement decision dependent on  $\theta_p$ , the person-specific substantive trait, and the extreme decision dependent on  $\eta_p$ , the person-specific ERS. Therefore, the ordinal category probability is obtained by:

$$p(X_{pi} = x_{pi}) = \left[ \frac{\exp(y_{1pi}(\theta_p - \beta_{1i}))}{1 + \exp(\theta_p - \beta_{1i})} \right] \left[ \frac{\exp(y_{2pi}(\eta_p - \beta_{2i}))}{1 + \exp(\eta_p - \beta_{2i})} \right], \quad (1)$$

where  $\beta_{hi}$  denotes the difficulty of pseudo-item  $h$  of item  $i$ . Note that this definition of only one pseudo-item describing both extreme decision nodes reflects the assumption of identical decision-making processes for extreme agreement and disagreement (i.e., directional invariance of extreme responding, see Jeon & De Boeck, 2019).

### 3.2. Multidimensional Node Parameterization

The traditional IRTree model with unidimensional nodes implies that each decision during the response selection is based on only one personal characteristic, either the substantive trait or a RS. However, as derived above, we rather assume that respondents consistently make a trait-based global decision to agree vs. disagree, but that the fine-grained decision in favor of the particular extreme or moderate category is guided by both trait-based and ERS-based processes, the composition of which is dependent on test-taking effort. In order to model this assumption, the extreme decision nodes can be extended by within-node multidimensionality (see Jeon & De Boeck, 2016; Meiser et al., 2019; von Davier & Khorramdel 2013) so that the respective pseudo-item responses are affected by both the trait and the ERS. In addition, the strict Rasch assumption of homogeneous item discrimination can be weakened by a 2PL parameterization with item-specific loadings of the person parameters so that influences of trait and ERS are not restricted to be constant throughout the questionnaire, but can vary depending on the item and its position within the questionnaire.

Figure 1 specifies this multidimensional 2PL parameterization of extreme responding, in which the item-specific response strategy is reflected by the loadings  $\alpha_i^{(\theta)}$  and  $\alpha_i^{(\eta)}$  of the substantive trait  $\theta$  and ERS  $\eta$ , respectively. Further note that the extreme pseudo-item is split between the categories of disagreement and agreement, as proposed by Meiser et al. (2019): If the decision of agreement is answered affirmatively ( $y_{1pi} = 1$ ), extreme agreement is modeled to be more likely under high trait levels and high ERS levels. For disagreeing responses ( $y_{1pi} = 0$ ), in contrast, the trait loadings are set to be negative, so that high trait levels increase the probability of moderate

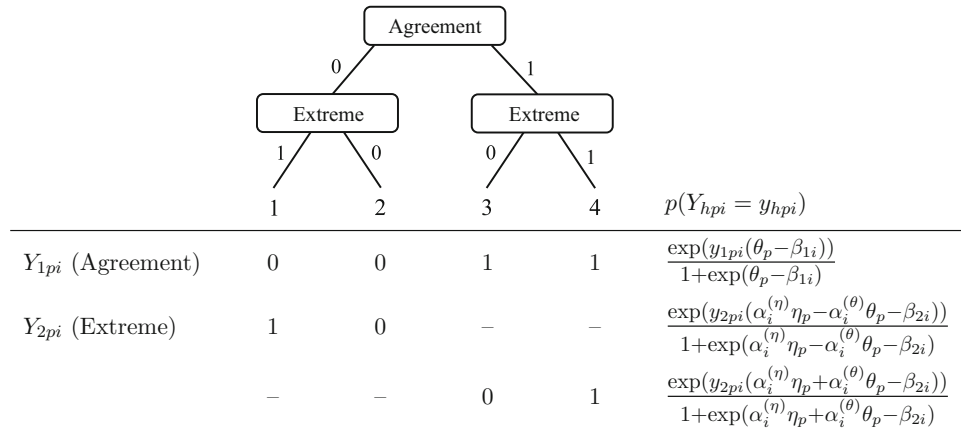


FIGURE 1.

Tree diagram, definition of pseudo-items, and multidimensional node probabilities for responses to four-point Likert-type items. Due to the conditional definition of extreme responding, one of the two pseudo-item variants is missing by design for each ordinal category, as indicated by '-'. The item-specific loadings are constrained with  $\alpha_i^{(\eta)} \geq 0$  and  $\alpha_i^{(\theta)} \geq 0$ .

(i.e., non-extreme) disagreement, whereas high ERS levels still make extreme disagreement more likely. Therefore, the ordinal category probability is obtained by

$$p(X_{pi} = x_{pi}) = \left[ \frac{\exp(y_{1pi}(\theta_p - \beta_{1i}))}{1 + \exp(\theta_p - \beta_{1i})} \right] \left[ \frac{\exp(y_{2pi}(\alpha_i^{(\eta)} \eta_p + \alpha_i^{(\theta)} \theta_p - \beta_{2i}))}{1 + \exp(\alpha_i^{(\eta)} \eta_p + \alpha_i^{(\theta)} \theta_p - \beta_{2i})} \right]^{y_{1pi}} \left[ \frac{\exp(y_{2pi}(\alpha_i^{(\eta)} \eta_p - \alpha_i^{(\theta)} \theta_p - \beta_{2i}))}{1 + \exp(\alpha_i^{(\eta)} \eta_p - \alpha_i^{(\theta)} \theta_p - \beta_{2i})} \right]^{1 - y_{1pi}}, \tag{2}$$

with  $\alpha_i^{(\eta)} \geq 0$  and  $\alpha_i^{(\theta)} \geq 0$ .

#### 4. The Dynamic Response Strategy Model

The novel dynamic response strategy model (DRSM) bases on the multidimensional IRTree parameterization and accounts for dynamic changes of response strategies over the course of the questionnaire by modeling the loadings of response processes as a function of item position. We use a modified form of the gradually changing function proposed by Wollack and Cohen (2004) and Goegebeur et al. (2008), which can capture linear as well as curvilinear relationships of a response process  $p$ , and is given by:

$$\alpha_i^{(p)} = \left( \gamma_1^{(p)} - \gamma_I^{(p)} \right) \left( 1 - \left( \frac{i - 1}{I - 1} \right)^{\lambda^{(p)}} \right) + \gamma_I^{(p)}, \tag{3}$$

with  $\gamma_1^{(p)} \geq 0$ ,  $\gamma_I^{(p)} \geq 0$ , and  $\lambda^{(p)} \geq 0$ . The parameters  $\gamma_1^{(p)}$  and  $\gamma_I^{(p)}$  are the loadings of process  $p$  of the first and last item, respectively. The actual dynamic change is captured by the slope, which is the difference between the last and first loadings ( $\gamma_I^{(p)} - \gamma_1^{(p)}$ ). Therefore, a positive slope reflects a dynamically increasing influence, a negative slope reflects a decreasing influence,

and a zero-slope trajectory reflects a non-dynamic, constant influence of the respective response process. In the further course of the article, we will frequently refer to the absolute slope, which accordingly describes the strength of the change, irrespective of the direction. The parameter  $\lambda^{(p)}$  determines the shape of the trajectory for process  $p$  over item position, which is linear for  $\lambda^{(p)} = 1$ , and curvilinear otherwise (see Fig. 2).

The proposed DRSM for dynamic response strategies of extreme decisions can be derived by inserting a dynamic loading trajectory described by Eq. 3 into each the trait loadings and ERS loadings in Eq. 2. Thereby, the process loadings of the DRSM are defined to be nonnegative across all items, which is a frequently made assumption in IRT modeling (e.g., Jin & Wang, 2014; Kim & Bolt, 2021; Meiser et al., 2019). We consider this a reasonable constraint also for the loading trajectories, since variations in test-taking effort should result in a varying degree of trait and RS involvement, that is, in a varying size of the loadings, whereas a change toward negative loadings would rather imply a qualitatively different effect of such latent personal characteristics on response selection (e.g., high trait levels would then be associated with low instead of high response categories). Nonetheless, the DRSM could likewise be specified without this constraint, by allowing  $\gamma_1^{(p)}$  and  $\gamma_I^{(p)}$  to vary freely, in order to put the underlying assumption to the test. The consideration of negative loadings could additionally be a useful extension when some items are inverted with regard to the substantive trait,<sup>1</sup> meaning that high trait levels are associated with endorsements of lower response categories. For such items, the DRSM could be adjusted by inverting the signs of the trait loadings in Eq. 2, so that the loadings of extreme agreement would be constrained negative and the loadings of extreme disagreement positive. The direction in which the trait influences the response selection (i.e., toward higher or lower categories) could thus be determined individually for each item, while the process loadings defined in Eq. 3 would reflect the strength with which a process is involved, without specifying the direction. Further note that the above parameterization of the DRSM refers to a fixed item order across respondents, as the same index  $i$  is used for the difficulty parameters and the response process loadings. An alternative approach would be the presentation of items in person-specific random order, for which the model can be adjusted accordingly, by defining item-dependent difficulties and position-dependent loadings.

Other reasonable modifications of the DRSM will be illustrated in this article, such as integrating dynamic influences of response processes not only into the two-dimensional decision nodes of extreme responding, but also into the unidimensional trait-based agreement decision (see Sect. 6). Further, IRTree models for response scales with more than four categories often include additional pseudo-items and additional RS, like decisions of moderate responding dependent on MRS, and such can likewise be modeled by the DRSM (see Sect. 7). Moreover, the DRSM as described above considers item position as the only predictor of response process loadings, thus implying a continuous response strategy with monotonically changing loadings. This is a theoretical model with an explicit focus on item position as *one* of possibly several factors influencing the impact of trait-based and RS-based processes within each item. However, in the context of the second simulation study, we derive a flexible extension of the DRSM, which still accounts for dynamic loading trajectories, but at the same time can capture further (random) item-specific variation of loadings.

<sup>1</sup>We thank an anonymous reviewer for pointing out the importance of distinguishing between positive and negative response process loadings.

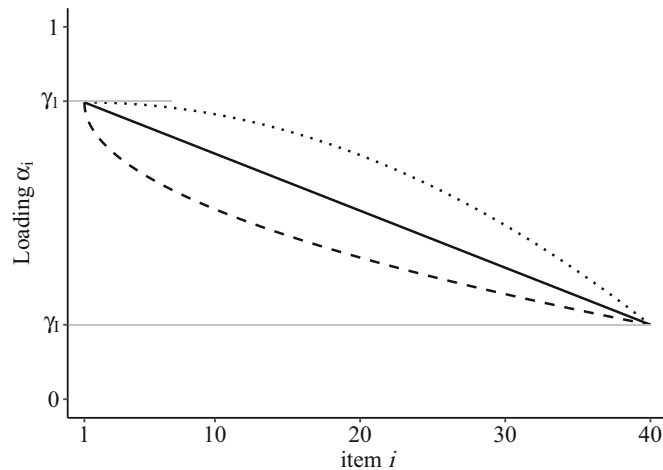


FIGURE 2.

Relationship of loadings  $\alpha_i$  and item position  $i$  for  $I = 40$  items with  $\gamma_1 = 0.8$ ,  $\gamma_I = 0.2$  and  $\lambda = 1$  (solid line),  $\lambda = 2$  (dotted line), and  $\lambda = 0.5$  (dashed line).

#### 4.1. Evaluating the Dynamic Response Strategy Model

Two simulation studies were conducted to systematically evaluate the proposed dynamic modeling approaches (DRSM and an extended version), and to provide answers to the following questions: Firstly, is the DRSM an appropriate cognitive explanatory model that reliably detects and quantifies dynamic influences of response processes in the data? Secondly, is the DRSM a beneficial psychometric measurement model that creates added value for the analysis of item response data over existing models? Both questions were investigated under ideal conditions, in which the data-generating model followed a continuous, model-implied response strategy (Sect. 5), and under more realistic conditions, in which additional random variation was added (Sect. 6). The proposed dynamic models were evaluated in relative comparison to IRTree models representing reasonable alternatives.

### 5. Simulation Study 1

In the first simulation study, we addressed dynamic and non-dynamic continuous response strategies, meaning that the trait and ERS loadings were constrained by continuous trajectories so that influences of response processes were only depend on the position of items within the questionnaire. For such scenarios, we examined the accuracy of the DRSM in recovering dynamic response strategies (i.e., sensitivity to detect changes in the impact of trait and ERS), as well as the risk of false-positive dynamics, which is finding such an effect if it is not present in the data (i.e., specificity of unveiling dynamic changes only in cases where they do exist). Further, we investigated the recovery of person and item parameters and the out-of-sample model fit in comparison with alternative IRTree models.

#### 5.1. Models of Continuous Response Strategies

The simulation study covered item response data under the assumption of continuous response strategies, and all applied models can be derived from the general model structure described in Fig. 1 and Eq. 2. The parameterization of the agreement pseudo-item was equal across models and corresponded to a solely trait-dependent unidimensional Rasch model. The two-dimensional

TABLE 1.  
Loading constraints of models with continuous response strategies used in simulation study 1.

Model	Trait loading $\alpha_i^{(\theta)}$	ERS loading $\alpha_i^{(\eta)}$
DRSM	$(\gamma_1^{(\theta)} - \gamma_I^{(\theta)}) \left( 1 - \left( \frac{i-1}{I-1} \right)^{\lambda^{(\theta)}} \right) + \gamma_I^{(\theta)}$	$(\gamma_1^{(\eta)} - \gamma_I^{(\eta)}) \left( 1 - \left( \frac{i-1}{I-1} \right)^{\lambda^{(\eta)}} \right) + \gamma_I^{(\eta)}$
Static	$\alpha^{(\theta)}$	$\alpha^{(\eta)}$
ERS	0	1
Ordinal	1	0

definition of the extreme pseudo-item with item-specific loadings of trait and ERS served as a superordinate framework, from which we derived unidimensional as well as two-dimensional special cases, as defined in Table 1.

Under the DRSM, the item-specific loadings were determined by the trait and ERS trajectories, which comprise the three parameters  $\gamma_1$ ,  $\gamma_I$ , and  $\lambda$  each. The static model excludes such a change over items, but rather assumes a single constant loading for each of the two processes. Besides those models with two-dimensional response strategies, also unidimensional models were derived. Their extreme decision nodes were Rasch parameterized and included only one of the person parameters with constant loading 1, that is the ERS  $\eta$  for the ERS model and the substantive trait  $\theta$  for the ordinal model. The loadings of the respective other parameter were set to 0. The ERS model is equivalent to the traditional IRTree described by Eq. 1 and the ordinal model corresponds to its counterpart as was proposed by Kim and Bolt (2021).

## 5.2. Data Generation

Using R (R Core Team, 2020), item response data were generated according to the four IRTree models with continuous response strategies described above. In the DRSM, the trait and ERS loading trajectories were systematically varied to cover a wide range of plausible dynamic response strategies (resulting in six model variants). We only simulated decreasing trait loadings and increasing ERS loadings, as these correspond to our theoretical consideration, and analogous models can be specified for opposite trajectories. The trait loading trajectories were generated with  $(\gamma_1; \gamma_I)$  set to (0.8; 0.2), (0.7; 0.3), (0.6; 0.4), and (0.5; 0.5), so that the absolute slopes were of size 0.6, 0.4, 0.2, and 0.0. Likewise, the ERS loading trajectories were set to (0.2; 0.8), (0.3; 0.7), (0.4; 0.6), and (0.5; 0.5).<sup>2</sup> We combined trait and ERS trajectories with different absolute slopes so that the response strategy change was either large (i.e., one process changed by 0.6, the other by 0.4), medium (0.4 and 0.2), or small (0.2 and 0.0). For each generated data set, both trajectories were generated with the same value of  $\lambda$ , set to 2 or 0.5, which we considered as reasonable values for a positively or negatively accelerated dynamic change, respectively. For data generation with the static model, two model variants were defined, which are the constant trait and ERS loadings  $(\alpha^{(\theta)}; \alpha^{(\eta)})$  set to (0.3; 0.7) and (0.7; 0.3). The two unidimensional models have fixed trait and ERS loadings and thus do not require to specify additional parameters.

For all model variants, 100 replications were conducted each for two sample sizes  $N$ , set to 500 and 1000, and with the two questionnaire lengths  $I$ , set to 20 and 40. Each data set consisted

<sup>2</sup>The center of all trait and ERS trajectories, that is  $(\gamma_1 + \gamma_I)/2$ , was set to 0.5. This value was chosen for both response processes because a prior study (Meiser et al., 2019) and our empirical application indicated (1) that trait-based and RS-based processes have about the same influence on two-dimensional IRTree decisions, and (2) that the loadings of the substantive trait are roughly twice as large in the broad agreement decision compared to fine-grained response selection (e.g., in extreme decision nodes).

of binary responses to the two pseudo-items of agreement and extreme responding under a certain model variant and was generated as follows: Firstly, the person parameters, that are  $N$  trait levels  $\theta_p$  and  $N$  ERS levels  $\eta_p$ , were generated to be uncorrelated and sampled from independent standard normal distributions. Likewise,  $2I$  pseudo-item difficulties  $\beta_{hi}$  were randomly drawn from a standard normal distribution. Then, person and item parameters were inserted into the respective equation of the model variant with its item-specific trait and ERS loadings. Lastly, for each person and each item, binary responses to the pseudo-items were randomly sampled according to the model-implied probabilities.

### 5.3. Model Estimation and Analysis

Each data generation step was followed by a model estimation step, in which all four models with continuous response strategy changes (see Table 1) were applied to the respective data set. In addition, also a 2PL model with freely estimated item-specific trait and ERS loadings was fitted (the agreement node was Rasch parameterized as in the other models). The 2PL model is not specifically targeted at continuous response strategies, but as all previously described continuous models are nested within it, it could be a flexible, universal alternative.

Bayesian parameter estimation was performed using the No-U-Turn Sampler (Hoffman & Gelman, 2014), a Markov chain Monte Carlo algorithm implemented in the software program Stan (Carpenter et al., 2017). R served as the interface to Stan along with the package CmdStanR (Gabry & Cešnovar, 2021). Four chains were run with each 1000 iterations and a warmup of 500 iterations. All estimated models reached convergence, indicated by values of the potential scale reduction factor  $\hat{R}$  less than 1.05. Note that all point estimates reported in the following sections are the expected a posteriori (EAP) estimates.

Priors were chosen according to recommendations in the Bayesian IRT literature (e.g., Luo & Jiao, 2018; Stan Development Team, 2020). The priors for  $\theta_p$  and  $\eta_p$  were set to standard normal distributions, and a normally distributed hierarchical prior was applied to the item difficulties  $\beta_{hi}$  with a *Cauchy*(0, 5) hyperprior for the mean and nonnegative *Cauchy*(0, 5) for the standard deviation. Weakly informative *LogNormal*(0, 2) priors were placed on (1)  $\gamma_1$  and  $\gamma_I$  of the DRSM trajectories, (2) the constant loadings of the static model, and (3) the item-specific trait and ERS loadings of the 2PL model ( $\alpha_i^{(\theta)}$ ,  $\alpha_i^{(\eta)}$ ). This ensured the convergence of the models even under conditions in which a response process did not contribute to the data generation, but needed to be estimated by the model (e.g., estimating the influence of the ERS for ordinal data). The shape  $\lambda$  of the DRSM was defined in the interval [0.25, 4] and given a *LogNormal*(-0.5, 1) prior. Again, this prior was chosen to ensure convergence, as only weak information is available to estimate  $\lambda$  if the difference of  $\gamma_1$  and  $\gamma_I$  is small. In an extreme case of a zero-slope trajectory,  $\lambda$  can take any value without having an effect on the loading trajectory. Further, the interval boundaries were chosen to account for the fact that  $\lambda$  and  $\frac{1}{\lambda}$  have symmetrical effects on the curvature of the trajectory, though in the opposite directions (see Fig. 2) and assured that the estimated shapes were within the range of plausible trajectories that we considered to be continuously and not abruptly changing.

### 5.4. Results

The DRSM and the four alternative models considered in the simulation study only differed in their constraints regarding the loadings of response processes. To give an overview of how the models behaved when fitted to item response data generated under such constraints, Fig. 3 illustrates the trait loading estimates provided by the different models for exemplary data sets. Even though these examples cannot summarize the entirety of simulations, beneficial characteristics of



the DRSM in comparison with alternative models become clear, which we will elaborate on in the following.<sup>3</sup>

*5.4.1. Sensitivity: Estimation of Dynamic Trajectories* The first aim of the simulation was to examine the sensitivity of the DRSM, and accordingly, to answer the question of whether it is suitable for detecting and quantifying dynamic trajectories of trait and ERS loadings. Therefore, we evaluated (1) the recovery of the slope ( $\gamma_I - \gamma_1$ ), which is probably the most informative dynamic measure, as it quantifies the change over the course of the questionnaire, (2) the recovery of the shape  $\lambda$ , and (3) the precision of estimates of the trajectory parameters  $\gamma_1$ ,  $\gamma_I$ , and  $\lambda$ .

Figure 4a, b summarizes the slope estimates of trait and ERS loadings by the DRSM and reveals a good recovery of both trajectories. Irrespective of sample size and questionnaire length, the means across simulation replications closely matched the respective true generated values. The good recovery of slopes is in line with consistently small posterior *SDs* of  $\gamma_1$  and  $\gamma_I$  (two bottom lines in Fig. 4c), meaning that they were estimated quite precisely. In contrast,  $\lambda$  estimates had high uncertainty and the posterior *SDs* were a multiple of those of the other two parameters. Note that the difference in posterior *SDs* between  $\lambda$  set to 0.5 and 2, as apparent in Fig. 4c, stems from the nonlinear relationship of  $\lambda$  and trajectory curvature, since the smaller  $\lambda$  is, the larger the change in the degree of curvature induced by slight changes in the parameter. In general, all three trajectory parameters were estimated more precisely the larger the size of the data set, determined by  $N$  and  $I$  (see Table A1). Moreover, the larger the absolute slope, the higher the precision of  $\lambda$  estimates, whereas estimates of  $\gamma_1$  and  $\gamma_I$  were not affected by the slope.

The link between slope and shape was already discussed with regard to the informative prior on  $\lambda$ , which was imposed to assure convergence in case of flat trajectories. Accordingly, the  $\lambda$  estimates of trajectories with small absolute slopes were highly influenced by the prior, which moved the parameter toward the value 1 (see Fig. 4d). The steeper the trajectories, the more information regarding the shape was provided by the data, the prior had less influence, and the  $\lambda$  estimates moved closer to the true values of 2 and 0.5, respectively. However, even for large slopes, the uncertainty of  $\lambda$  estimates was high so that EAP estimates should not be interpreted in a substantial manner without considering the uncertainty (e.g., the rough classification as positively or negatively accelerated trajectory is possible if the credible interval (CI) does not contain the value 1). Altogether, the analysis of sensitivity demonstrated that the DRSM is an appropriate cognitive model for investigating dynamic response strategies; it was highly suitable to detect systematic changes of response process influences, to assess the magnitude of such changes with high precision, and to roughly inform about the shape of the trajectory, if the DRSM itself was used to generate the data.

*5.4.2. Specificity: Estimation of Non-dynamic Trajectories* Furthermore, also the specificity was to be examined, that is, whether the model accurately detected the absence of dynamic changes. To answer this question, we evaluated simulation conditions in which the DRSM was fitted to non-dynamic data generated by the unidimensional models and the static model, which all have zero-slope trajectories. Indeed, the estimated slopes by the DRSM were all very close to 0 and had low variance (see Table A2), demonstrating that the model consistently detected the absence of dynamic changes (also see the illustration of estimated trajectories in exemplary zero-slope data sets in Fig. 3). The more parsimonious models were successfully mimicked, meaning that the estimated parameters by the DRSM reflected the restrictions of models nested within it. Therefore, the DRSM is a suitable cognitive model also for data with non-dynamic response strategies.

<sup>3</sup>The detailed results of all replications conducted in the simulation study can be found on the Open Science Framework repository: <https://osf.io/kc8ve/>.

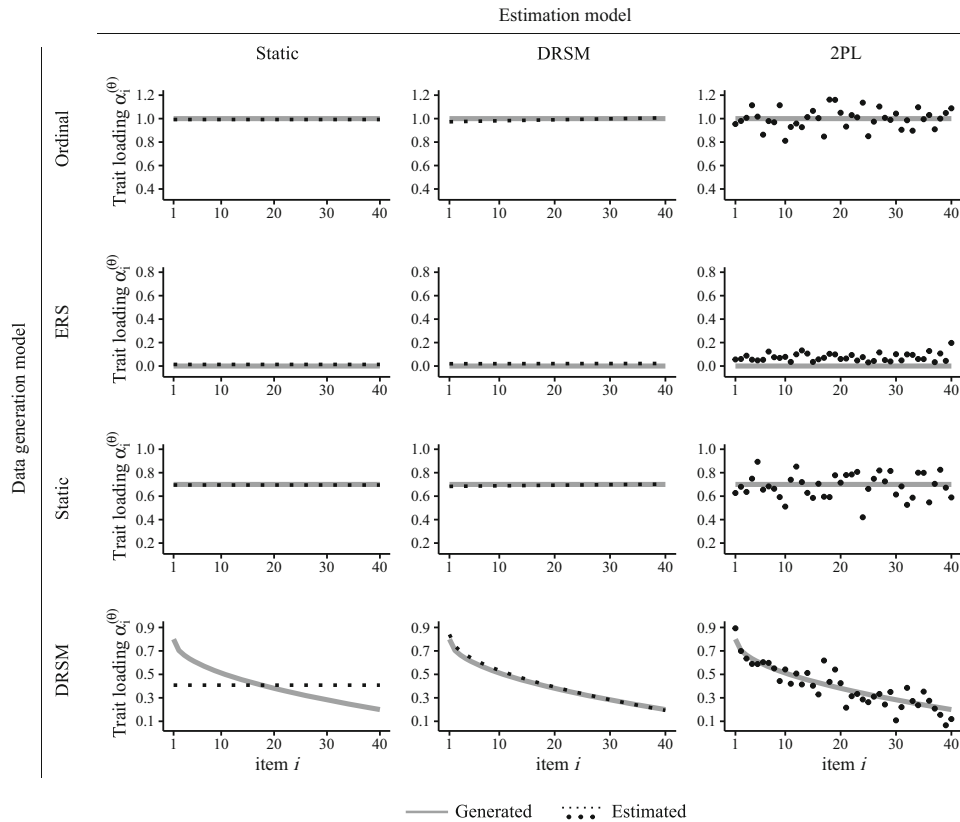


FIGURE 3.

Trait loading estimates of the static model, DRSM, and 2PL model to exemplary data sets generated by the four models of continuous response strategies used in simulation study 1. Trait loadings of the ordinal and ERS model are not shown, as they are not estimated, but fixed at 1 and 0, respectively.

**5.4.3. Parameter Recovery** Besides investigating the adequacy of the DRSM to accurately describe different response strategies, we aimed to examine its quality as a psychometric model. To this end, the recovery of person parameters ( $\theta_p$  and  $\eta_p$ ) and item parameters ( $\beta_{hi}$  and  $\alpha_i$ ) was compared across models, measured by root mean square error (RMSE). There were only minor differences in parameter recovery between conditions with different sample sizes or questionnaire lengths, except that the overall levels of RMSEs were smaller, the larger the data set (see Fig. A1). The results for conditions with  $N = 1000$  and  $I = 40$  are illustrated in Fig. 5. In general, the models with two-dimensional extreme decision nodes (static model, DRSM, and 2PL model) yielded considerably smaller errors compared to the unidimensional models (ordinal and ERS). The unidimensional models showed good parameter recovery only for data sets generated by the respective model itself but performed poorly for data generation with all other models.

The two-dimensional models, in contrast, were nearly equally well suited to recover the parameters of all kinds of data sets, with the exception of the response process loadings. Here, the 2PL model revealed comparably large errors, which is in line with the finding that its freely estimated loadings do not perfectly mimic the underlying dynamic or non-dynamic continuous trajectories, but rather scatter around them (see Fig. 3). The static model and DRSM recovered all parameters of models nested within them well, though unsurprisingly, the static model showed larger errors for the loadings of the DRSM. This shows that the DRSM was the model with the overall best parameter recovery and that it did not pose a risk of increased errors in case of misspecifications, but successfully mimicked lower-parameterized models.



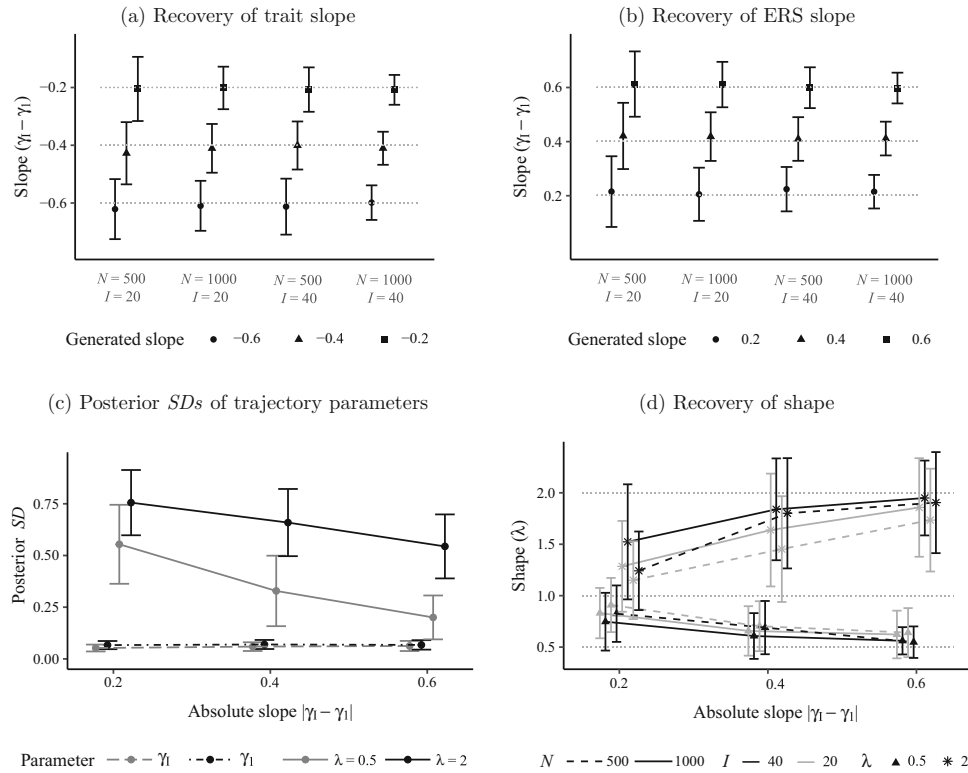


FIGURE 4.

Estimates and precision of trajectory parameters by the DRSM for continuous dynamic data in simulation study 1. Error bars represent the *SDs* of estimates across simulation replications.

**5.4.4. Model Fit** The five models were further compared with regard to their fit, for which we calculated the out-of-sample prediction accuracy by an approximation of leave-one-out cross-validation (LOO; Vehtari et al., 2017). LOO is a fully Bayesian information criterion, which has been shown to outperform alternative methods like Akaike’s information criterion (AIC; Akaike, 1974) or the deviance information criterion (DIC; Spiegelhalter et al., 2002) in IRT model selection (Luo & Al-Harbi, 2017). Table 2 lists the average LOO information criterion values (small values indicate better fit), as well as the proportion of simulation replications in which the respective model was the best one in predicting the data.

Across all conditions, the respective data-generating model itself provided the best out-of-sample fit in the majority of replications and entailed the smallest average LOO values. The DRSM was almost always selected as the best-fitting model for data sets with dynamic response strategies of medium or large size. For small response strategy changes (i.e., one process changed with absolute slope 0.2, the other was constant with slope 0.0), the DRSM was still the best-fitting model, though also the static model often predicted the data well. This condition of small dynamic changes was the only one in which we found substantial differences across sample sizes and questionnaire lengths: For  $N = 500$  and  $I = 20$ , the DRSM itself was selected in only 50 % of replications, whereas it was selected in 92 % for  $N = 1000$  and  $I = 40$ . Thus, only if limited data was available, the zero-slope trajectories of the static model gave a good approximation of dynamic loadings. The larger the data set, the more evident the advantage of the more complex DRSM and its additionally estimated parameters was.

For data sets generated by the unidimensional models, not only the respective model itself, but also the static model and DRSM predicted the data well, as can be seen by average LOO values

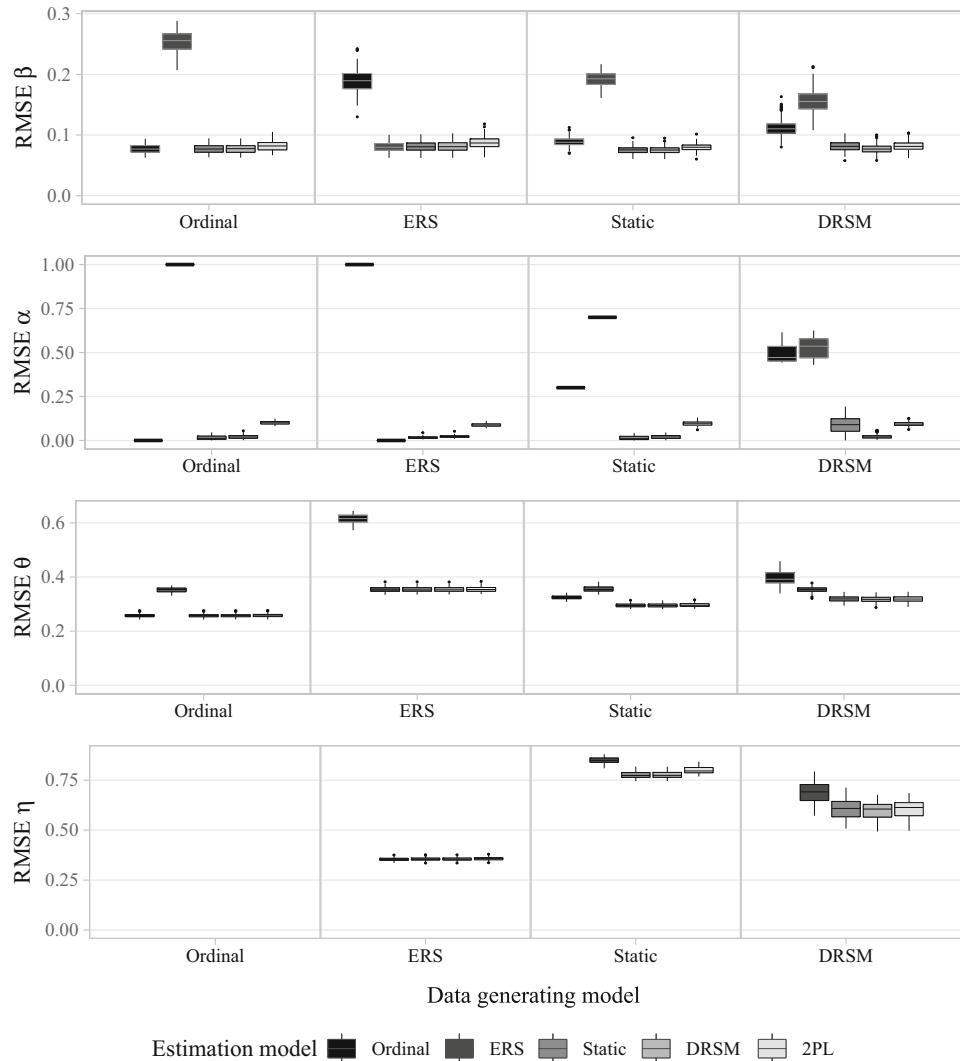


FIGURE 5.

RMSEs of estimated person and item parameters for continuous data in simulation study 1. The boxplots summarize the results for the simulation condition with  $N = 1000$  and  $I = 40$ . For data generation with the static model, only the condition with  $\alpha_i^{(\theta)} = 0.7; \alpha_i^{(\eta)} = 0.3$  is shown. RMSEs of  $\eta$  estimates for data generated with or estimated by the ordinal model are missing, as the model does not incorporate an ERS influence.

quite close to the values of the data-generating model. This can be explained by the fact that both two-dimensional models closely mimicked the lower-parameterized models and did not overfit despite their greater flexibility (see the good recovery of person and item parameters in Figs. 5 and 3). The same holds true for static data, which was mostly best predicted by the static model itself, but for which the DRSM also revealed small LOO values. In contrast, the 2PL model was hardly ever selected as the best-fitting model, and the average LOO values were considerably larger than those of the DRSM. The analysis of the parameter recovery suggests that these indicators of overfitting and reduced generalizability of the 2PL model are mainly due to the comparably poor recovery of loading estimates.

The DRSM thus offered the best compromise between flexibility on the one hand, and generalizability on the other hand, if dynamic or non-dynamic continuous response strategies are present in the data. The model not only proved to be a valuable cognitive model of response

TABLE 2.  
Model comparisons by LOO out-of-sample prediction accuracy for continuous data in simulation study 1.

Data generation		Average LOO information criterion					Proportion of replications in favor of				
Model	Abs. slopes	ORD	ERS	Static	DRSM	2PL	ORD	ERS	Static	DRSM	2PL
ORD	0/0	<b>87,419</b>	92,860	87,422	87,423	87,492	<b>0.68</b>	0	0.22	0.10	0
ERS	0/0	96,649	<b>88,135</b>	88,137	88,139	88,212	0	<b>0.58</b>	0.28	0.14	0
Static	0/0	92,071	91,416	<b>89,751</b>	89,754	89,853	0	0	<b>0.74</b>	0.26	0
DRSM	0.2/0.0	91,885	91,465	90,036	<b>90,019</b>	90,125	0	0	0.25	<b>0.75</b>	0
	0.4/0.2	92,053	91,755	90,264	<b>90,163</b>	90,266	0	0	0.01	<b>0.99</b>	0
	0.6/0.4	91,872	91,665	90,107	<b>89,841</b>	89,942	0	0	0	<b>1.00</b>	0

The LOO values and proportions in bold indicate the overall best-fitting model in the respective data generation condition.

The average LOO information criterion values include the replications with  $N = 1000$  and  $I = 40$ . The other conditions yielded comparable patterns.

strategies with high sensitivity (i.e., dynamic changes of response process influences over items are reliably captured) and high specificity (i.e., dynamic changes are not falsely revealed)—it is further a beneficial psychometric measurement model, which provides added value in terms of parameter recovery and models selection.

## 6. Simulation Study 2

In order to further investigate the benefits of the DRSM under real-world conditions, the second simulation study addressed non-continuous dynamic response strategies. They are characterized by a general trend of loadings over the course of the questionnaire, but, in contrast to continuous strategies, allow for item-specific deviations from the trajectories. Such loading patterns are probably more frequently encountered in empirical data than strictly continuous trajectories, because even though previous research indicated that item position is a crucial factor influencing the impact of response processes, additional item-specific variation of loadings could arise; for instance, due to the items' levels of abstractness or complexity (e.g., positively vs. negatively worded items, grammatical or linguistic complexity). Such additional variance between items does hardly affect the overall dynamic response strategy change across item positions and, as a result, should not limit the validity of the DRSM as a cognitive model. Thus, the first aim of the simulation study is to put this assumption to the test and to analyze the performance of the DRSM in detecting dynamic changes in the presence of additional random variation of loadings. However, even if general changes in the response behavior can be detected, the DRSM would simplify the true, underlying data-generating processes. Therefore, we propose a more flexible extension, the F-DRSM, which can capture the hypothesized non-continuous dynamic response strategies in addition to systematic underlying trajectories. Thus, the second aim of the simulation study is to evaluate the F-DRSM and to examine its psychometric properties.

### 6.1. The Flexible Dynamic Response Strategy Model

The F-DRSM can be seen as a combination of the DRSM and 2PL model: The item-specific loadings are the sum of a systematic component, which is defined by a continuous trajectory, and unsystematic, random noise. Fixing this unsystematic component for each item to 0 would yield the DRSM, whereas omitting the systematic trajectory component would result in the 2PL model. In the F-DRSM, the random noise is assumed to stem from a common normal distribution across

all items, in which the mean is fixed to zero, and the standard deviation indicates the strength of the deviation from the trajectory. The F-DRSM loadings of a response process  $p$  are defined by:

$$\begin{aligned} \alpha_i^{(p)} &\sim \text{Normal}(\mu_i^{(p)}, \sigma^{(p)}), \\ \mu_i^{(p)} &= (\gamma_1^{(p)} - \gamma_I^{(p)}) \left( 1 - \left( \frac{i-1}{I-1} \right)^{\lambda^{(p)}} \right) + \gamma_I^{(p)}, \\ \sigma^{(p)} &\sim \text{Cauchy}(0, 5). \end{aligned} \tag{4}$$

The F-DRSM provides estimates for (1) the item-specific loadings  $\alpha_i^{(p)}$ , (2) the underlying dynamic trajectory, and (3) the standard deviation  $\sigma^{(p)}$ , with which the loadings scatter around the trajectory. Thereby, the item-specific loadings are allowed to deviate from the respective values predicted by the common trajectory ( $\mu_i^{(p)}$ ) but are at the same time shrunken to this mean. The trajectory can thus be seen as a prior for the free loadings, which is not fixed but estimated from the data. Whether this prior is rather informative or uninformative depends on the data: If almost all item-specific loadings of a process closely fit a trajectory, the remaining loadings are strongly shrunken to this trajectory. In contrast, if the data suggest that the loadings largely scatter and do not form a trajectory, the individual loading estimates are hardly affected by the trajectory estimate, and the F-DRSM converges to the standard 2PL model. As the extent of unsystematic variation of loadings does not have to be determined a priori, but emerges as an estimate from the model, hypotheses regarding the strength of the continuous trend can be tested. Moreover, the variance of loadings does not have to remain unsystematic but can be explained by further predictors.

### 6.2. Data Generation and Model Estimation

All generated data sets of the second simulation study contained item responses of  $N = 1000$  respondents to  $I = 40$  items on a four-point scale under the model assumptions of the F-DRSM. In contrast to the first simulation study, not only the two-dimensional IRTree nodes of extreme responding but also the unidimensional agreement node was given a 2PL parameterization with item-specific trait loadings, so that the ordinal category probability was given by:

$$\begin{aligned} p(X_{pi} = x_{pi}) &= \left[ \frac{\exp(y_{1pi}(\alpha_{1i}^{(\theta)}\theta_p - \beta_{1i}))}{1 + \exp(\alpha_{1i}^{(\theta)}\theta_p - \beta_{1i})} \right] \left[ \frac{\exp(y_{2pi}(\alpha_i^{(\eta)}\eta_p + \alpha_{2i}^{(\theta)}\theta_p - \beta_{2i}))}{1 + \exp(\alpha_i^{(\eta)}\eta_p + \alpha_{2i}^{(\theta)}\theta_p - \beta_{2i})} \right]^{y_{1pi}} \\ &\quad \left[ \frac{\exp(y_{2pi}(\alpha_i^{(\eta)}\eta_p - \alpha_{2i}^{(\theta)}\theta_p - \beta_{2i}))}{1 + \exp(\alpha_i^{(\eta)}\eta_p - \alpha_{2i}^{(\theta)}\theta_p - \beta_{2i})} \right]^{1-y_{1pi}}. \end{aligned} \tag{5}$$

Thus, three noisy trajectories were defined, that are the trait loadings of agreement ( $\alpha_{1i}^{(\theta)}$ ), the trait loadings of extreme responding ( $\alpha_{2i}^{(\theta)}$ ), and the ERS loadings of extreme responding ( $\alpha_i^{(\eta)}$ ). The absolute slopes of these three trajectories were varied (0.0, 0.2, 0.4, 0.6), whereby all trait trajectories decreased over items and the ERS trajectories increased. The standard deviation of the unsystematic noise was set to either 0.1 or 0.2 for all trajectories. All further settings were chosen as in the first study. Examples of the resulting non-continuous dynamic loadings for different underlying trajectories are given in Fig. 6. We generated 100 data sets for each model variant and fitted the DRSM, the F-DRSM, and the 2PL model, using the same Bayesian parameter estimation and priors as in the first study.

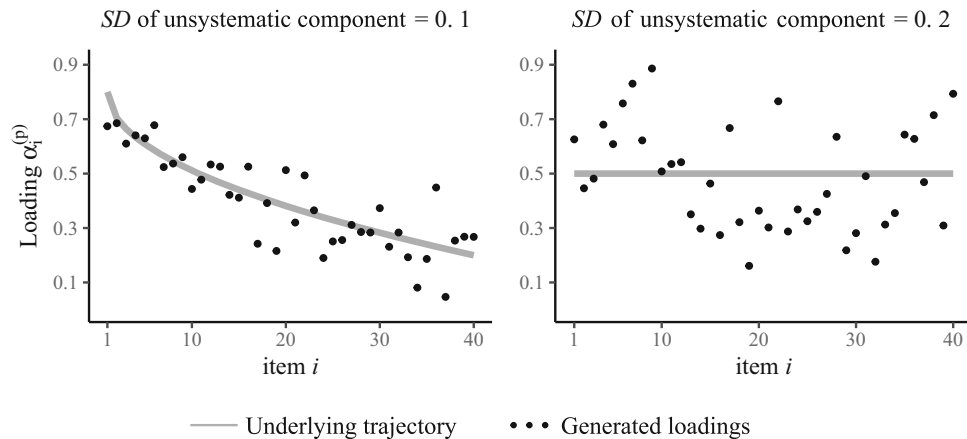


FIGURE 6.  
Examples of randomly generated loadings under the F-DRSM.

### 6.3. Results

**6.3.1. Slope Estimates** To evaluate the utility of the two dynamic models as cognitive models for describing systematic changes in response strategies, we analyzed the recovery of the slopes of the three response process trajectories. In addition, also the slope recovery of the 2PL model was investigated, which does not provide estimates of trajectories inherently, so that we fitted dynamic functions through the freely estimated loadings in a post hoc analysis. In contrast to the F-DRSM, the estimation of the loadings in the 2PL model is independent of the estimation of the trajectories.

Notably, the DRSM, F-DRSM, and 2PL model recovered the slopes equally well, irrespective of the size of the slope and irrespective of the  $SD$  of the unsystematic component, demonstrating that dynamic as well as non-dynamic response processes were successfully detected (see Table A3 for the comparison of slope estimates by the three models). Moreover, the data set-specific slope estimates of DRSM, F-DRSM, and 2PL model hardly differed from each other and were highly correlated, which underlines that the models drew almost identical conclusions regarding the response strategy changes. This suggests that, for the sole sake of determining the size of strategy changes, there is no practical difference between (1) taking a continuous dynamic trajectory as a direct estimate of individual loadings, (2) using it as a prior that shrinks the loadings, or (3) fitting the trajectory after estimating the loadings freely. However, the models largely differed in the uncertainty with which the slopes were estimated, as the DRSM yielded considerably higher precision of trajectory parameter estimates and had the smallest posterior  $SD$ s of slopes. Thus, the DRSM enables more specific conclusions to be drawn about the extent of response strategy changes (e.g., whether the change is different from 0, or whether there are differences between groups of items or persons), so that it should be preferred over the other models for response behavior analyses.

**6.3.2. Model Fit and Parameter Recovery** Furthermore, the three models also differed in their suitability as psychometric models: Model comparisons by LOO out-of-sample prediction accuracy clearly showed the benefits of the F-DRSM, which provided the smallest LOO values and was chosen as the best-fitting model in all replications (see Table 3). Even under conditions with large unsystematic  $SD$  of 0.2, in which the item-specific loadings largely scatter so that trajectories are hardly recognizable (see Fig. 6), the F-DRSM was advantageous over the unrestricted estimation by the 2PL model. Unsurprisingly, the DRSM could not predict the data well, since it cannot properly capture the non-continuous patterns of loadings.

TABLE 3.  
Model comparisons by LOO out-of-sample prediction accuracy for non-continuous data in simulation study 2.

Data generation		Average LOO information criterion			Proportion of replications in favor of		
Abs.	<i>SD</i>	DRSM	F-DRSM	2PL	DRSM	F-DRSM	2PL
slope	0.0	90,230	<b>90,140</b>	90,210	0	<b>1.00</b>	0
	0.2	90,365	<b>89,836</b>	89,864	0	<b>1.00</b>	0
0.2	0.1	89,969	<b>89,883</b>	89,952	0	<b>1.00</b>	0
	0.2	90,509	<b>89,987</b>	90,015	0	<b>1.00</b>	0
0.4	0.1	89,926	<b>89,840</b>	89,904	0	<b>1.00</b>	0
	0.2	89,814	<b>89,328</b>	89,354	0	<b>1.00</b>	0
0.6	0.1	89,605	<b>89,521</b>	89,582	0	<b>1.00</b>	0
	0.2	89,638	<b>89,162</b>	89,184	0	<b>1.00</b>	0

The LOO values and proportions in bold indicate the overall best-fitting model in the respective data generation condition.

Likewise, the analysis of parameter recovery demonstrated the superiority of the F-DRSM, which yielded the smallest errors across person and item parameters (see Fig. 7). As was the case in the first simulation study, also the 2PL model provided a good recovery, except for slightly higher RMSEs of loadings. The DRSM showed comparably high errors for the item parameters but still recovered the person-specific substantive trait levels as accurately as the higher-parameterized models. This suggests that the model adjusted the item difficulties in a way that they counteracted the deviations of individual loadings from a continuous trajectory. In line with this, the smaller the unsystematic *SD*, the smaller the disadvantage of the DRSM compared to the more flexible models.

Overall, the second simulation study clearly showed the benefits of modeling dynamic response strategies under real-world conditions and revealed that both the DRSM and the F-DRSM are models with high utility, though for different kinds of research goals. On the one hand, the DRSM accurately reflected the magnitude of response strategy changes over the items of a questionnaire, and although the two alternative models produced almost identical results at the level of point estimates, the DRSM had the advantage of more precise estimates. Therefore, we recommend using the DRSM as a cognitive explanatory model if the goal is to analyze data sets with a focus on investigating the respondents' behavior. Further, the recovery of person-specific trait levels was hardly affected by model choice, making the DRSM a parsimonious and convenient model for controlling substantive trait estimates for dynamic RS effects. Nevertheless, the DRSM inevitably simplifies the true loading patterns whenever strategy changes are not perfectly continuous. Therefore, the more flexible F-DRSM is the preferable model for investigating influences of response processes on the level of individual items, as it can accurately capture random fluctuations of response strategies in addition to the underlying continuous trend. Further, it was even better suited than a 2PL model in terms of model selection and parameter recovery, so the F-DRSM has proven to be a beneficial psychometric model for the analysis of item response data under influence of response style effects.

### 7. Empirical Application

To demonstrate the advantages of modeling response strategy changes in real data, we applied the proposed dynamic models as well as the other models used in the simulation studies to an empirical data set taken from Johnson (2014). The data consists of item responses to the

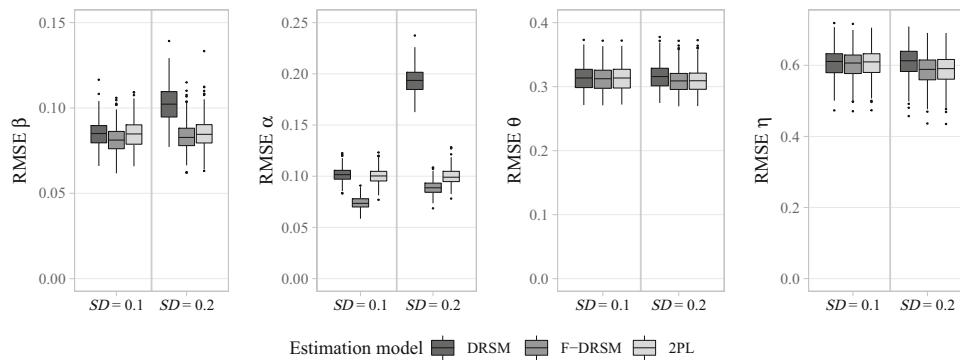


FIGURE 7.

RMSEs of estimated person and item parameters for non-continuous data in simulation study 2.

IPIP-NEO-120 with 120 items on five broad personality domains (Neuroticism, Extraversion, Conscientiousness, Agreeableness, and Openness to Experience), and we randomly selected 1000 participants of the large online sample from the American population with complete data. The response scale of the IPIP-NEO-120 comprises five rating categories, which is why we extended the tree structure of Fig. 1 by an additional node of moderate vs. non-moderate responding (e.g., Böckenholt, 2012; Böckenholt & Meiser, 2017; Khorramdel & von Davier, 2014). The extended tree of five-point rating scales is depicted in Fig. 8.

The decision of moderate responding was modeled to be dependent on the respondent's mid-point RS (MRS)  $\kappa_p$ , weighted with an item-specific loading  $\alpha_i^{(\kappa)}$ , and on the item difficulty  $\beta_{0i}$  of the additional pseudo-item ( $h = 0$ ). As was done in the second simulation study, the trait loadings of the agreement node ( $\alpha_{1i}^{(\theta)}$ ) were not fixed but estimated so that our theoretical assumption of dynamically changing response strategies predominantly occurring in fine-grained decisions could be empirically tested. Thus, all four response processes (MRS-based moderate responding, trait-based agreement, trait-based extreme responding, and ERS-based extreme responding) had independent item-specific loadings, which were either fixed (ordinal and ERS model), or estimated according to the model-specific constraints. Further, we let the variances of person parameters be estimated unless a standard normal prior was needed for model identification (i.e., one of the five traits and both RS for the static model and DRSM; all traits and both RS for the F-DRSM and 2PL model). We used the same Bayesian model estimation as in the simulation studies, with four chains, 500 warmup iterations, and 1000 post-warmup iterations to assure convergence (all  $\hat{R} < 1.05$ ).

For the analysis of dynamic response strategy changes, we examined the estimates of the DRSM, as the simulations suggested that it provides the most precise trajectory estimates. As hypothesized, the estimated slopes of MRS loadings  $\alpha_i^{(\kappa)}$  and ERS loadings  $\alpha_i^{(\eta)}$  were of substantial size and were both significantly larger than 0, meaning that the response strategy changed toward more RS involvement (see Table 4). Moreover, the influence of the substantive trait on fine-grained extreme decisions decreased, as indicated by a negative slope of loadings  $\alpha_{2i}^{(\theta)}$ , which is in line with the idea of RS-based processes taking over from trait-based responding in two-dimensional pseudo-items. This is original evidence that fine-grained decisions—and in particular, the RS-based processes—are highly dependent on the item position, suggesting an increase in fatigue and satisficing over the course of the questionnaire. Contrary to our assumption, we found that the trait loadings of the broad agreement decision ( $\alpha_{1i}^{(\theta)}$ ) also decreased over time. However, as will be argued below, this is probably due to other item characteristics than the position within the questionnaire. Furthermore, the  $\lambda$  estimates were 0.48 (MRS), 0.33 (ERS), 0.67 (trait-based agreement), and 0.59 (trait-based extreme responding), though only the 95 % CIs of the two RS



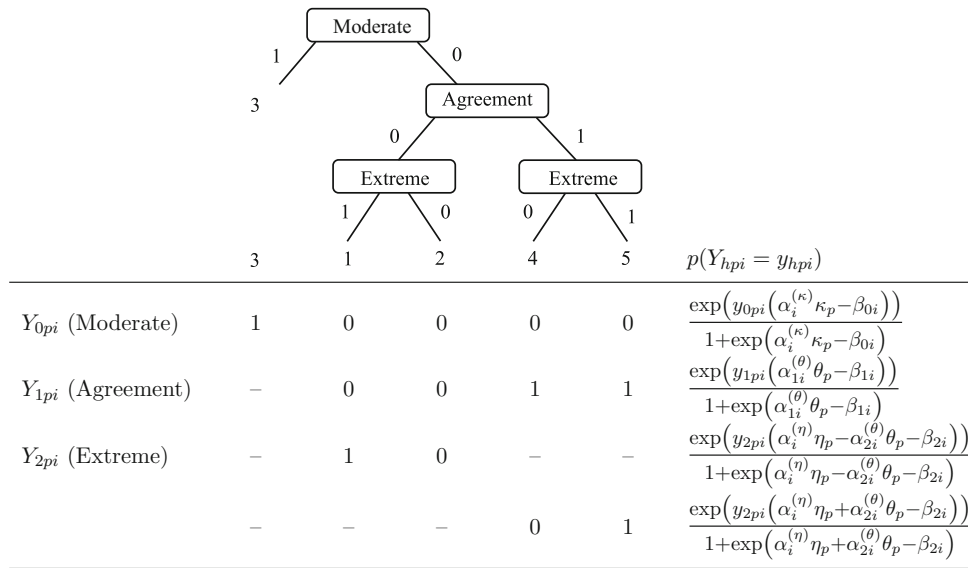


FIGURE 8.

Tree diagram, definition of pseudo-items, and node probabilities for responses to five-point Likert-type items. Pseudo-items that are missing by design are marked with ‘-’.

trajectories did not include 1. This indicates that the response strategy change was most dominant in earlier item positions and is decelerated at the end of the questionnaire (also see Fig. 9).

The models were further compared by the LOO information criterion, which demonstrated that the F-DRSM provided the best fit (see Table 4). As was the case in the second simulation study, shrinking loadings to a dynamic trajectory was advantageous over freely estimating them or constraining them to a continuous function. The shrinkage effect is clearly visible in Fig. 9, which displays the estimated item-specific loadings by the 2PL model as well as the loadings and trajectories by the F-DRSM. Notably, the MRS and ERS loadings of the F-DRSM closely scatter around the respective estimated dynamic trajectories and the estimated *SDs* are 0.15 and 0.12, respectively. In contrast, the trait loadings of agreement and extreme responding scatter much stronger and the estimated *SDs* are 0.74 and 0.41, respectively.<sup>4</sup> This indicates that the item position is a crucial determinant of the impact of RS-based response processes, whereas trait-based responding is greatly affected by other factors (e.g., other item characteristics or item content).

To give an example of how such additional factors can be investigated, we fitted an additional model with separate dynamic trajectories for positively and negatively worded items. For three of the four response processes, there were hardly any differences between the loading trajectories of positively and negatively worded items.<sup>5</sup> Only trait-based agreement ( $\alpha_{1i}^{(\theta)}$ ) differed largely between conditions: Positively worded items had high trait loadings throughout the questionnaire and the slope was not significantly different from zero ( $\gamma_1 = 1.524$ ;  $\gamma_I = 1.345$ ), whereas negatively worded items started with significantly smaller loadings, which then increased over items ( $\gamma_1 = 0.990$ ;  $\gamma_I = 1.265$ ). In the later part of the questionnaire, the conditions did not differ anymore (at 5 % error level). This suggests that it took the respondents some practice to process negatively worded items as accurately as positively worded ones. Further, negatively worded items

<sup>4</sup>The second simulation study revealed accurate recovery of the *SD* of the unsystematic component. Across replications with low *SD* of 0.1:  $M = 0.103$ ,  $SD = 0.023$ ; with high *SD* of 0.2:  $M = 0.201$ ,  $SD = 0.030$ .

<sup>5</sup>The only significant differences at 5 % error level were in  $\gamma_I$  and slope of the trait loadings of extreme responding ( $\alpha_{2i}^{(\theta)}$ ), which indicated that loadings of positively worded items had a stronger decline.



TABLE 4.  
Model fit and slope estimates for the empirical data set.

Model	LOO	Estimated slope and 95 %-CI			
		$\alpha_i^{(\kappa)}$	$\alpha_i^{(\eta)}$	$\alpha_{1i}^{(\theta)}$	$\alpha_{2i}^{(\theta)}$
Ordinal	311,417	0	0	0	0
ERS	305,271	0	0	0	0
Static	300,667	0	0	0	0
DRSM	300,433	0.31 [0.20, 0.42]	0.54 [0.43, 0.66]	-0.28 [-0.43, -0.16]	-0.12 [-0.22, -0.03]
F-DRSM	296,649	0.30 [0.14, 0.49]	0.54 [0.35, 0.72]	-0.34 [-0.92, 0.24]	-0.08 [-0.37, 0.28]
2PL	296,718	0.31 [0.13, 0.50]	0.57 [0.36, 0.79]	-0.38 [-1.02, 0.19]	-0.02 [-0.35, 0.31]

$\alpha_i^{(\kappa)}$  = loadings of MRS-based moderate responding;  $\alpha_i^{(\eta)}$  = loadings of ERS-based extreme responding;  $\alpha_{1i}^{(\theta)}$  = loadings of trait-based agreement;  $\alpha_{2i}^{(\theta)}$  = loadings of trait-based extreme responding.

occur more frequently in the later part of the IPIP-NEO-120, which would lead to artifacts such as the unexpected negative slope of trait-based agreement in the models not accounting for item wording (see  $\alpha_{1i}^{(\theta)}$  in Table 4).

Taken together, our empirical application gave a first indication that there are qualitatively different mechanisms behind (1) response process of fine-grained decisions, which seem to be susceptible to response strategy changes toward heuristic responding (e.g., due to loss of motivation or fatigue) and (2) broad agreement decisions, which appear not to be affected by satisficing and reduced test-taking effort. However, more research focusing on response strategies as a psychological, cognitive construct is needed, in order to better understand how respondents arrive at their decisions, why they change their behavior over time, and how covariates (e.g., item wording) can explain different strategies.

## 8. Discussion

The present research introduced the dynamic response strategy model (DRSM) as well as a more flexible extension (F-DRSM) and demonstrated how these models overcome the previous limitation of response style (RS) modeling, being that systematically changing influences of response processes over the items of a questionnaire were not accounted for. The new approaches address such dynamic response behaviors by modeling item position-dependent loadings of response processes (e.g., trait-based or RS-based response selection) in unidimensional and two-dimensional IRTree decision nodes. These loadings are assumed to follow continuous or non-continuous trajectories, which can either be dynamically changing with a linear or curvilinear shape, reflecting a response strategy change throughout the questionnaire, or be static ones, reflecting a constant response behavior. While the DRSM is an idealized and theory-driven model of strictly continuously changing loadings, the F-DRSM is adapted to more realistic, non-continuous settings, in which the item-specific loadings scatter around an underlying trajectory with normally distributed noise.

Simulation studies were conducted to compare the dynamic models and reasonable alternative models in terms of their suitability as cognitive explanatory models (i.e., requiring accurate quantification of dynamic changes of response processes and correct identification of non-dynamic, constant response strategies) as well as psychometric measurement models (i.e., requiring good recovery of person and item parameters and model fit). The DRSM turned out to be a capable cognitive model, as it reliably captured dynamically changing influences of response processes,

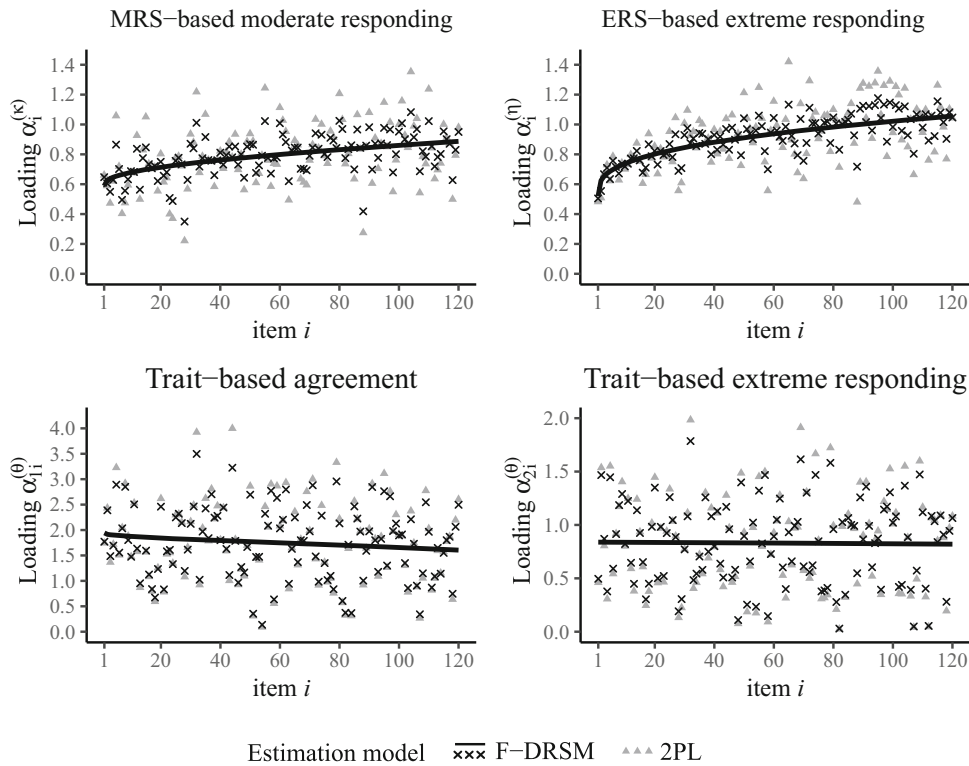


FIGURE 9.  
Loading estimates by the F-DRSM and 2PL model to the empirical data set.

irrespective of whether the data followed a continuous or non-continuous response strategy. Moreover, it did not pose the danger of misinterpretations due to falsely identified dynamics in the case of constant response strategies. Although perfectly continuous trajectories of response process loadings are most likely a simplification of real-world settings, the DRSM is a parsimonious formalization of the underlying general trend of the respondents' behavior. Regardless of whether the model can capture all characteristics of a data set well, it proved to be a simple but appropriate model for investigating response strategies and changes of those over items. As such, it can be used to investigate how respondents arrive at their decisions, to uncover increasing satisficing or reduced test-taking effort, to evaluate questionnaires with regard to their cognitive load and burden, and to compare subgroups of respondents or items. The empirical application demonstrated that it is worthwhile to address such research questions, as the findings provided new insights on factors influencing the impact of different response processes on response selection, which eventually affect the data quality. Knowing about such factors, like item wording, can inform and improve test construction. Furthermore, constraining loadings to a trajectory is easy to implement, also outside Bayesian estimation methods, and requires only minor adjustments of the traditionally applied IRTree models and only few additional parameters to estimate. Therefore, the DRSM is a convenient model in applied fields of research to analyze personal characteristics, attitudes, or beliefs, and to control substantive trait estimates for dynamic RS effect.

Nevertheless, under realistic conditions of non-continuous response strategy changes, the constraint of continuous trajectories inevitably impairs the accurate representation of the true item-specific response behavior. Accordingly, the simulations showed that the more flexible models with item-specific influences of response process were superior in terms of psychometric properties. Notably, the F-DRSM with the underlying assumption of the response strategy being

a composition of a systematic item position-dependent component and unsystematic noise was preferable not only to the continuous DRSM, but also to the unconstrained estimation of loadings by the 2PL model; both in simulated data and in the empirical application. Whenever the influences of response processes follow some sort of systematic over items, the F-DRSM effectively uses the information on this general trend for the estimation, making it a successful analysis model with good psychometric properties. The F-DRSM additionally comes with the advantage that the extent of unsystematic variance is estimated by the model so that the relative importance of the item position for the item-specific response strategy can be determined. For example, the empirical application not only showed that the influences of different response processes changed to different degrees over items, but also that the unsystematic variance greatly differed between trait-based and RS-based responding. The F-DRSM, therefore, allows to examine response strategies beyond the effect of the item position and opens up new possibilities for investigating the roles of different response processes in item responding.

Overall, the two new dynamic models provide a surplus value to previous RS modeling because they facilitate the investigation of systematic heterogeneity of response processes across items of a questionnaire, which goes beyond existing approaches on discrete heterogeneity across measurement situations (e.g., Ames & Leventhal, 2021; Colombi et al., 2021; Weijters et al., 2010) or classes of respondents (e.g., Gollwitzer et al., 2005; Khorramdel et al., 2019; Meiser & Machunsky, 2008; Tijmstra et al., 2018). Thereby, our proposed models are a valuable first step toward transforming the heterogeneity of response processes over items into systematic variance, and toward answering the question of how respondents arrive at their judgments and decisions.

### *8.1. Limitations and Future Directions*

A limitation of the proposed models is that we defined the loading trajectories at the group level, which was based on the assumption that the factors leading to a change in the response strategy across items (e.g., loss of motivation) apply to all respondents to a similar extent. The estimated loading trajectories, however, can only represent the average response behavior and do not allow a differentiated assessment of individual response processes. In contrast, the mixture IRTree model by Kim and Bolt (2021) facilitates to distinguish groups of respondents by their response strategy, though under the strict assumption that all decisions during response selection are unidimensional and based either on the substantive trait or on the ERS, without considering a combination of processes, and without accommodating systematic changes. An integration of both approaches in the sense of a mixture DRSM (or F-DRSM) seems promising, as it may allow identifying classes of respondents with a dynamically changing response strategy, with a constant response strategy, and also respondents with only one of the two processes involved (which then follow the ERS or ordinal model). Such a model could be used in future studies to examine the heterogeneity of response processes between persons and simultaneously account for within-person changes across the items of a questionnaire.

A limitation of the conducted simulation studies is the chosen range of dynamic scenarios, which only partially covers all conceivable empirical response strategy changes. For instance, the simulated variation of the unsystematic component of the F-DRSM was lower than those we found for some processes in the empirical application. However, the process loadings in the simulation were centered around 0.5, while they were substantially larger in the application, which could naturally lead to a higher variance of the noise. Nonetheless, the conclusions drawn from the simulation study are limited to small or medium unsystematic variation, while the suitability of the F-DRSM for larger variations of loadings was suggested by empirical findings.

Further investigations might also be needed with regard to the shape of loading trajectories, as the proposed models are based on monotonous functions of loadings. If the response strategy changes were non-monotonic, for example, because respondents need to warm up to the task in

the first part of a questionnaire, but lose motivation later on, U-shaped functions might be more appropriate. Even more complex loading patterns could result if respondents repeatedly alternate between responding with high and low effort, or if they start changing their strategy at a certain point within the questionnaire rather than from the very first item. Such complex patterns could in principle be investigated using the DRSM by replacing the curvilinear trajectories with other functions that match those hypotheses, although this would be challenging for model estimation and would likely require much larger data sets to achieve satisfactory precision of the results.

Also, the loadings might not only be associated with the time respondents spent on the test, which is captured by the item position, but further be dependent on the item content. For instance, in multidimensional questionnaires (such as the IPIP-NEO-120 used in the empirical application), the respondents' interest or expertise might differ between trait dimensions, leading to a correlation of response process loadings with the measured dimensions. Though we have not addressed such dependencies, they could be empirically tested and accounted for by estimating dimension-specific trajectories, or by explaining the unsystematic variation of loadings in the F-DRSM by trait dimensions.

In addition, the present research investigated dynamic response strategies exclusively in the framework of IRTree models. However, our modeling approach with focus on response process loadings could likewise be integrated in other IRT model classes for which extensions for response styles were developed, such as the generalized partial credit model (Muraki, 1992) or the graded response model (Samejima, 1969). Notwithstanding such conceivable extensions, with the DRSM and F-DRSM, we have presented two of several ways to approach dynamic response processes in item responding, and leave further adaptations, developments, and generalizations to future research.

### Acknowledgments

The authors thank the anonymous reviewers for their helpful comments on earlier drafts of this article.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This research was funded by the Deutsche Forschungsgemeinschaft (DFG) Grant 2277, Research Training Group Statistical Modeling in Psychology (SMiP).

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### References

- Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, 42(3), 957–970. <https://doi.org/10.1016/j.ssresearch.2013.01.002>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. Wiley.

- Ames, A. J., & Leventhal, B. C. (2021). Modeling changes in response style with longitudinal IRTree models. *Multivariate Behavioral Research*. <https://doi.org/10.1080/00273171.2021.1920361>
- Andersen, H., & Mayerl, J. (2017). Social desirability and undesirability effects on survey response latencies. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 135(1), 68–89. <https://doi.org/10.1177/0759106317710858>
- 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(6), 1235–1245. <https://doi.org/10.1016/j.paid.2005.10.018>
- Baumgartner, H., & Steenkamp, J.-B.E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, 17(4), 665–678. <https://doi.org/10.1037/a0028111>
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, 22(1), 69–83. <https://doi.org/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 and Statistical Psychology*, 70(1), 159–181. <https://doi.org/10.1111/bmsp.12086>
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, 71(5), 814–833. <https://doi.org/10.1177/0013164410388411>
- Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021a). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24(4), 718–738. <https://doi.org/10.1177/1094428120947794>
- Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021b). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*. <https://doi.org/10.1177/109442812111056520>
- Callegaro, M., Yang, Y., Bholra, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 103(1), 5–25. <https://doi.org/10.1177/075910630910300103>
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1–32. <https://doi.org/10.18637/jss.v076.i01>
- Colombi, R., Giordano, S., & Kateri, M. (2021). Hidden markov models for longitudinal rating data with dynamic response styles. <https://arxiv.org/pdf/2111.13370>
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, 48(1), 1–28. <https://doi.org/10.18637/jss.v048.e01>
- DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: A literature review. *Quality and Quantity*, 52(4), 1523–1559. <https://doi.org/10.1007/s11135-017-0533-4>
- Deutskens, E., de Ruyter, K., Wetzels, M., & Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters*, 15(1), 21–36. <https://doi.org/10.1023/B:MARK.0000021968.86465.00>
- Gabry, J., & Češnovar, R. (2021). Cmdstanr: R interface to CmdStan.
- Galesic, M. (2006). Dropouts on the web: Effects of interest and burden experienced during an online survey. *Journal of Official Statistics*, 22(2), 313–328.
- Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, 73(2), 349–360. <https://doi.org/10.1093/poq/nfp031>
- Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, 73(1), 65–87. <https://doi.org/10.1007/s11336-007-9031-2>
- Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, 17(1), 56–69. <https://doi.org/10.1037/1040-3590.17.1.56>
- Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, 25(5), 560–576. <https://doi.org/10.1037/met0000249>
- Henninger, M., & Plieninger, H. (2020). Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement. *Assessment*, 28(5), 1301–1319. <https://doi.org/10.1177/1073191119900003>
- Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, 45(4), 549–559. <https://doi.org/10.1086/268687>
- Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(47), 1593–1623.
- Huang, H.-Y. (2020). A mixture IRTree model for performance decline and nonignorable missing data. *Educational and Psychological Measurement*, 80(6), 1168–1195. <https://doi.org/10.1177/0013164420914711>
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, 48(3), 1070–1085. <https://doi.org/10.3758/s13428-015-0631-y>
- Jeon, M., & De Boeck, P. (2019). Evaluation on types of invariance in studying extreme response bias with an IRTree approach. *British Journal of Mathematical and Statistical Psychology*, 72(3), 517–537. <https://doi.org/10.1111/bmsp.12182>



- Jin, K.-Y., & Wang, W.-C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, 51(2), 178–200. <https://doi.org/10.1111/jedm.12041>
- Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51, 78–89. <https://doi.org/10.1016/j.jrp.2014.05.003>
- Kahn, R. L., & Cannell, C. F. (1957). *The dynamics of interviewing: Theory, technique, and cases*. Wiley.
- Kelava, A., & Brandt, H. (2019). A nonlinear dynamic latent class structural equation model. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(4), 509–528. <https://doi.org/10.1080/10705511.2018.1555692>
- 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(2), 161–177. <https://doi.org/10.1080/00273171.2013.866536>
- Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, 72(3), 538–559. <https://doi.org/10.1111/bmsp.12179>
- Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, 81(1), 131–154. <https://doi.org/10.1177/0013164420913915>
- Knowles, E. S., & Condon, C. A. (1999). Why people say “yes”: A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, 77(2), 379–386. <https://doi.org/10.1037/0022-3514.77.2.379>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, 50(1), 537–567. <https://doi.org/10.1146/annurev.psych.50.1.537>
- List, M. K., Robitzsch, A., Lüdtke, O., Köller, O., & Nagy, G. (2017). Performance decline in low-stakes educational assessments: Different mixture modeling approaches. *Large-scale Assessments in Education*. <https://doi.org/10.1186/s40536-017-0049-3>
- Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review*, 36(1), 116–124. <https://doi.org/10.1177/0894439317695581>
- Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling*, 59(2), 183–205.
- Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*. <https://doi.org/10.1177/0013164417693666>
- Marcus, B., Bosnjak, M., Lindner, S., Pilischenko, S., & Schütz, A. (2007). Compensating for low topic interest and long surveys: A field experiment on nonresponse in web surveys. *Social Science Computer Review*, 25(3), 372–383. <https://doi.org/10.1177/0894439307297606>
- Meiser, T., & Machunsky, M. (2008). The personal structure of personal need for structure. *European Journal of Psychological Assessment*, 24(1), 27–34. <https://doi.org/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 and Statistical Psychology*, 72(3), 501–516. <https://doi.org/10.1111/bmsp.12158>
- Messick, S. (1991). Psychology and methodology of response styles. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 161–200). Lawrence Erlbaum Associates.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- 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). Academic Press. <https://doi.org/10.1016/B978-0-12-590241-0.50006-X>
- Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, 53(5), 633–654. <https://doi.org/10.1080/00273171.2018.1469966>
- Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, 74(5), 875–899. <https://doi.org/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(5), 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- 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(1), 539–569. <https://doi.org/10.1146/annurev-psych-120710-100452>
- R Core Team. (2020). R: A language and environment for statistical computing. <https://www.R-project.org/>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1), 1–97. <https://doi.org/10.1007/BF03372160>
- Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, 81(4), 1118–1141. <https://doi.org/10.1007/s11336-015-9476-7>
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639. <https://doi.org/10.1111/1467-9868.00353>
- Stan Development Team. (2020). Stan modeling language users guide (2.26). <https://mc-stan.org>

- Suh, Y., Cho, S.-J., & Wollack, J. A. (2012). A comparison of item calibration procedures in the presence of test speededness. *Journal of Educational Measurement*, 49(3), 285–311. <https://doi.org/10.1111/j.1745-3984.2012.00176.x>
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, 38(5), 522–547. <https://doi.org/10.3102/1076998613481500>
- Tijmstra, J., & Bolsinova, M. (in press). Modeling within- and between-person differences in the use of the middle category in Likert scales. *Applied Psychological Measurement*. <https://research.tilburguniversity.edu/en/publications/modeling-within-and-between-person-differences-in-the-use-of-the->
- Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, 50(6), 2325–2344. <https://doi.org/10.3758/s13428-017-0997-0>
- Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge University Press.
- Ullitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, 87(2), 593–619. <https://doi.org/10.1007/s11336-021-09817-7>
- 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(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>
- von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 463–487). Springer. [https://doi.org/10.1007/978-1-4614-9348-8\\_30](https://doi.org/10.1007/978-1-4614-9348-8_30)
- von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). Springer. [https://doi.org/10.1007/978-0-387-49839-3\\_6](https://doi.org/10.1007/978-0-387-49839-3_6)
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, 15(1), 96–110. <https://doi.org/10.1037/a0018721>
- Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*, 33(5), 352–364. <https://doi.org/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(3), 279–291. <https://doi.org/10.1177/1073191115583714>
- Wollack, J. A., & Cohen, A. S. (2004). *A model for simulating speeded test data [Conference presentation]*. San Diego: Annual meeting of the American Educational Research Association.
- 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*, 22(1), 51–68. <https://doi.org/10.1002/acp.1331>
- Zettler, I., Lang, J. W. B., Hülshager, 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(4), 461–472. <https://doi.org/10.1111/jopy.12172>
- Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, 8(2), 127–135. <https://doi.org/10.18148/srm/2014.v8i2.5453>

Manuscript Received: 15 FEB 2022

Published Online Date: 6 FEB 2023





**Co-Occurring Dominance and Ideal Point Processes:  
A General IRTree Framework for Multidimensional Item Responding**


Viola Merhof and Thorsten Meiser

University of Mannheim

January 22, 2024

**Author Note**

Viola Merhof  <https://orcid.org/0000-0002-1328-0000>

Thorsten Meiser  <https://orcid.org/0000-0001-6004-9787>

Additional materials are available on OSF:

[https://osf.io/yu4gx/?view\\_only=50fc21d10d52414aaeece310d680fc0e](https://osf.io/yu4gx/?view_only=50fc21d10d52414aaeece310d680fc0e)

The data sets used for reanalyses in the empirical applications are made available by the original authors: <https://www.icpsr.umich.edu/web/HMCA/studies/6647> and <https://osf.io/gqb4y/>.

This research was funded by the Deutsche Forschungsgemeinschaft (DFG) grant 2277, Research Training Group Statistical Modeling in Psychology (SMiP).

This work was partly performed on the computational resource bwUniCluster funded by the Ministry of Science, Research and the Arts Baden-Württemberg and the Universities of the State of Baden-Württemberg, Germany, within the framework program bwHPC.

Correspondence should be made to Viola Merhof, Department of Psychology, University of Mannheim, L 13 15, D-68161 Mannheim, Germany. Email: [merhof@uni-mannheim.de](mailto:merhof@uni-mannheim.de)

**Abstract**

Responding to rating scale items is a multidimensional process, since not only the substantive trait being measured, but also additional personal characteristics can affect the respondents' category choices. A flexible model class for analyzing such multidimensional responses are IRTree models, in which rating responses are decomposed into a sequence of sub-decisions. Different response processes can be involved in item responding both sequentially across those sub-decisions and as co-occurring processes within sub-decisions. In the previous literature, modeling co-occurring processes has been exclusively limited to dominance models, where higher trait levels are associated with higher expected scores. However, some response processes may rather follow an ideal point rationale, where the expected score depends on the proximity of a person's trait level to the item's location. Therefore, we propose a new multidimensional IRT model of co-occurring dominance and ideal point processes (DI-MIRT model) as a flexible framework for parameterizing IRTree sub-decisions with multiple dominance processes, multiple ideal point processes, and combinations of both. The DI-MIRT parameterization opens up new application areas for the IRTree model class and allows the specification of a wide range of theoretical assumptions regarding the cognitive processing of item responding. A simulation study shows that IRTree models with DI-MIRT parameterization provide excellent parameter recovery and accurately reflect co-occurring dominance and ideal point processes. In addition, a clear advantage over traditional IRTree models with purely sequential processes is demonstrated. Two application examples from the field of response style analysis highlight the benefits of the general IRTree framework under real-world conditions.

*Keywords:* IRTree models, ideal point models, dominance models, multidimensional IRT, response styles

Likert-type rating scales are widely used to assess personality, attitudes, or beliefs via self-reports, and they are omnipresent in research and applied fields of psychology and social sciences. A popular item response theory (IRT) approach for analyzing such rating data are item response tree (IRTree) models (Böckenholt, 2012; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012; Jeon & De Boeck, 2016), which have been proven to be a broadly applicable model class offering high flexibility with regard to investigating response processes underlying respondents' judgments and decisions.

A key characteristic of IRTree models is their multidimensional nature with the underlying assumption that multiple response processes are sequentially<sup>1</sup> involved in the selection of response categories. This property arises from the decomposition of the ordinal rating responses into a sequence of pseudo-items, which represent the sub-decisions assumed to be taken by the respondents during response selection. For example, respondents may first decide on whether to agree or disagree with an item, and subsequently make fine-grained decisions among the available agreement or disagreement categories. Such sub-decisions are typically assumed to be binary judgments, though the pseudo-items can likewise be defined as ordinal judgments with three or more options (see Meiser et al., 2019, and Figure 1). The pseudo-items are modeled by separate IRT models, and by assigning different latent traits to the sub-decision, their effects on the response selection can be disentangled. Thus, IRTree models can capture multidimensional response processes, even though making use of unidimensional IRT modeling for the individual pseudo-items.

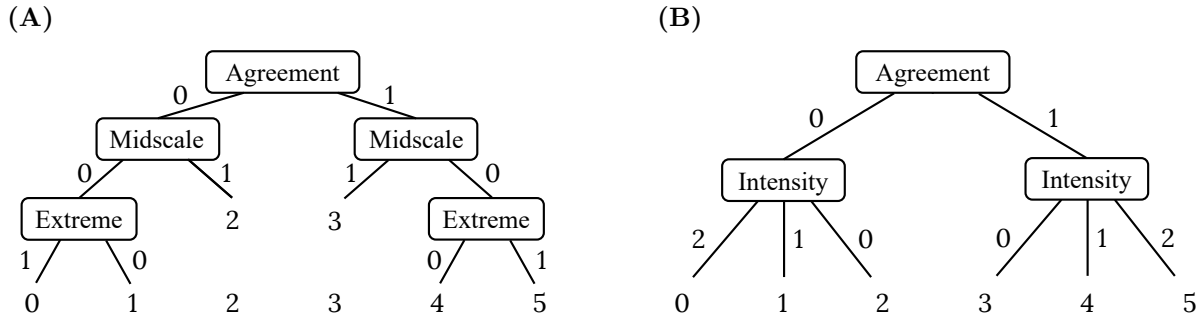
A typical aim of using IRTree models is to separate the effects of substantive traits from those of response styles (RS) – individual preferences for specific response categories of rating scales irrespective of item content (for an overview, see Van Vaerenbergh & Thomas, 2013). For instance, some respondents may prefer categories located in the middle of the response scale (midscale response style; MRS), whereas others may rather prefer clear-cut responses and tend to select the extreme categories (extreme response style; ERS). Such different usages of the response scale can systematically distort the estimation of individual substantive trait

---

<sup>1</sup>The term *sequential* in the context of IRTree models refers to the logical sequence and conditionality of response processes, which does not necessarily imply a temporal sequence.

**Figure 1**

*Tree Diagrams for Decomposing Responses to Six-Point Rating Items Into Sub-Decisions*



*Note.* A: Decomposition into three sub-decisions by five binary pseudo-items. Adapted from Böckenholt (2017). B: Decomposition into two sub-decisions by binary and ordinal (three-step) pseudo-items. Adapted from Meiser et al. (2019).

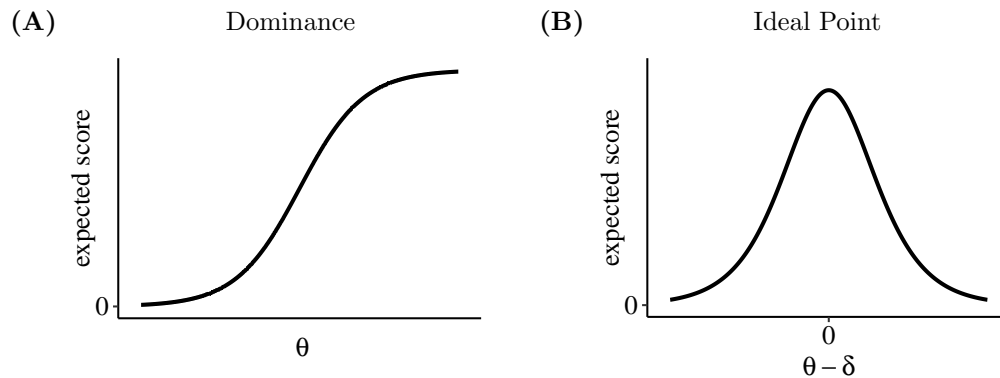
levels, group means, and correlations among multiple traits, so that RS must be controlled for to obtain valid measurements (Alwin, 2007; Baumgartner & Steenkamp, 2001). Commonly used IRTree models for the analysis of RS define agreement decisions as dependent on the substantive trait levels of respondents, whereas more fine-grained decisions are modeled to be based on individual RS, like the judgment to give extreme versus non-extreme responses guided by ERS, or the judgment to select the neutral middle category guided by MRS (e.g., Böckenholt, 2017; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013). However, even though IRTree models are mostly used for RS analysis, they are flexible to incorporate any kind of person-specific influences on the selection of individual response categories (e.g., socially desirable responding) by defining the pseudo-items correspondingly.

In contrast to this flexibility of IRTree models with regard to including various latent traits in the pseudo-items, little attention has so far been devoted to their flexibility in terms of modeling both monotonous and non-monotonous effects of such traits on the response selection. This property is described by the item response function (IRF), which defines how each value of the latent trait continuum maps to the expected score of an item. For ordinal item responses  $Y \in \{0, \dots, K\}$  on a rating scale with  $K+1$  categories, the IRF of trait  $\theta$  is given by

$$\text{IRF}(\theta) = \sum_{y=0}^K y \cdot p(Y = y | \theta). \quad (1)$$

Thus, the IRF defines the expected value of  $Y$  for a given trait level  $\theta$  depending on the category-specific probabilities that are specified by the IRT model (e.g., the generalized partial credit model; GPCM; Muraki, 1992).

IRT models can be grouped into two classes based on their IRFs: Dominance and ideal point models (Coombs, 1964). They go back to Likert (1932) and Thurstone (1928), respectively, and both have a long history in the psychometric literature. Typically, IRT decision processes are assumed to follow the dominance rationale, meaning that a higher trait level is modeled to result in a higher expected score of the respective pseudo-item (see Figure 2A). As suggested by the term *dominance*, respondents overcome an item if their trait level values exceed the item's level of difficulty. For instance, the probability to agree with an item, which states that environmental protection is an important issue, increases with higher levels of environmental awareness of the respondents. Dominance models have monotonically increasing IRFs and frequently applied members of this class are the models of the Rasch family; for example, the Rasch model (Rasch, 1960) or 2PL model (Birnbaum, 1968) for binary items, or the GPCM for ordinal items. An alternative assumption is captured by *ideal point* models, in which the relationship between the expected score and the latent trait is unimodal and non-monotonic (see Figure 2B). The expected score is highest if a respondent's trait level, which is called their ideal point, matches the item's location, and decreases with larger distances. The more the trait levels deviate from the item location in an upward or downward manner, the stronger respondents will dismiss the item content from above or from below, respectively. For example, respondents with moderate environmental awareness may agree with the statement that the current environmental regulations are adequate, whereas respondents who prefer either stricter or less strict regulations disagree, though for different reasons. As the IRFs are symmetrical about the item location, only the proximity of a respondent and the item, not the direction of a deviation, is relevant for the response selection. Several IRT models for dichotomous and polytomous items have been developed for the ideal point rationale, like the hyperbolic

**Figure 2***Item Response Functions Under the Dominance and Ideal Point Assumption*

cosine model (Andrich, 1995; Andrich & Luo, 1993) or the generalized graded unfolding model (Roberts et al., 2000).

Although rating scale items are mostly constructed under and analyzed by dominance approaches, there is compelling evidence that ideal point models often better describe item responding of non-cognitive constructs, and thus, should be considered when analyzing self-reported data (e.g., Liu & Wang, 2016; Roberts & Laughlin, 1996; van Schuur & Kiers, 1994; for an overview, see Drasgow et al., 2010). Nevertheless, compared to dominance IRT modeling, there is little research on multidimensional response processes so far, and most ideal point models treat item responses as solely dependent on the substantive trait to be measured, while ignoring possible other influences. This poses a threat to the validity of ideal point models whenever RS or other additional response processes are involved in item responding (Liu & Wang, 2019).

Multidimensional processes should, therefore, be investigated not only for dominance but also for ideal point items. However, this can require integrating response processes with different IRFs into multidimensional models, for instance, if RS are to be considered. Such are by definition dominance processes since higher RS levels reflect stronger preferences for certain response categories. A straightforward way to include RS into the analysis of ideal point items are IRTree models, as the pseudo-items are parameterized independently of each other, and thus, processes of different IRFs can be defined by existing unidimensional IRT models of the

dominance and ideal point rationale. Indeed, Jin et al. (2022) demonstrated the advantages of such an IRTree model, in which an ideal point model was applied to a trait-based sub-decision and a dominance model to an ERS-based sub-decision. In two application examples of attitudinal questionnaires, the authors showed that their model fitted the data better than both an ideal point model ignoring RS, and than classical dominance IRTree models accounting for RS.

This example nicely illustrates that the decomposition of multidimensional item responses into unidimensional decision processes with different kinds of IRFs provides high flexibility while keeping the modeling complexity low. Still, this advantage comes at the cost of the simplistic assumption that each cognitive processing step during response selection is based on one response process at a time (e.g., either the substantive trait or a RS). However, multiple response processes may not only contribute to item responding sequentially, but may also occur simultaneously on the level of sub-decisions. Such co-occurring processes can likewise be integrated into IRTree models by replacing the traditionally used unidimensional pseudo-items with multidimensional IRT (MIRT) models (see Jeon & De Boeck, 2016; Meiser et al., 2019; von Davier & Khorramdel, 2013). In RS analysis, for example, Meiser et al. (2019) showed that the selection of more or less extreme response categories was not only dependent on the individual ERS, but further influenced by the substantive trait levels of respondents. Such multidimensional parameterizations of pseudo-items allow the investigation of more complex and presumably more realistic hypotheses about the cognitive processing during item responding, and they were shown to be preferable over unidimensional ones with regard to psychometric properties (Merhof & Meiser, 2023).

Nevertheless, multidimensional pseudo-items have so far been exclusively applied to combinations of dominance processes. Dominance MIRT models can be derived from unidimensional ones by extending a single latent trait to a linear combination of multiple traits, and thus reflect the assumption that several processes contribute to the response selection in a cumulative, additive way (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011; Falk & Cai, 2016; Henninger & Meiser, 2020; Jin & Wang, 2014). Ideal point processes, in contrast, must not be considered

additive to other processes, as this would counteract the proximity concept.<sup>2</sup> Therefore, modeling co-occurring response processes under the ideal point assumption is less straightforward and there exist only few models that address this challenge, all of which focus on modeling RS in addition to trait-based responding to ideal point items. For instance, the approaches by G. Luo (1998) and Wang et al. (2013) implicitly account for RS in ideal point items by defining random category thresholds which vary across persons. Javaras and Ripley (2007) and Liu and Wang (2019) also use random thresholds, though specify such explicitly as a linear combination of different RS. However, none of the models treats RS as independent, stand-alone response processes, but all rather assume that they can only occur in the presence of another trait-based process, as they are defined as person-specific shifts of trait-based responding. In contrast, our understanding of item responding in the framework of IRTree models is that trait-based and RS-based responding are distinct processes, which can make both individual and combined contributions to the sub-decisions during response selection. Further, the models are only adapted to the co-occurrence of an ideal point trait and dominance RS, and do not generalize to other types of response processes with any IRF. None of the models provides a general formulation that consistently connects multiple response processes independent of dominance and ideal point assumptions.

Therefore, the aim of the present article is to provide a general IRTree framework which is independent of the choice of dominance or ideal point modeling, and in which multiple response processes can be involved in the response selection (a) sequentially across pseudo-items, and (b) as co-occurring processes within pseudo-items. While sequential multidimensionality can be implemented using existing IRT modeling, we propose a new approach for co-occurring response processes, in which multiple dominance processes, multiple ideal point processes, as well as a combination of both are modeled in a consistent manner. The new MIRT model of co-occurring

---

<sup>2</sup>Nevertheless, the assumption of cumulative effects can also be reasonable in ideal point modeling under certain circumstances: For instance, Cui (2008) proposed a multidimensional model for repeated measurements, in which the person-specific latent trait at a given time point  $t$  was modeled as the sum of the trait at the baseline time point plus the change from baseline to time  $t$ . Since both of the trait factors are on the same latent scale, their contributions to the response selection can be considered to be additive. In contrast, the present article rather addresses scenarios of multidimensional item responding in which different response processes relate to different constructs, so that their contributions cannot be aggregated in a cumulative way.



dominance and ideal point processes (DI-MIRT) is based on the divide-by-total framework for ordinal item responses (see Thissen & Steinberg, 1986). It can be used independently of IRTree models, though in this article we focus on that model class and demonstrate that a DI-MIRT parameterization can specifically benefit IRTree pseudo-items: The new DI-MIRT model can not only be used for modeling multidimensional response processes in ideal point items (see section Response Style Analysis in Ideal Point Items), but also for including ideal point processes into dominance items (e.g., when modeling the selection of midscale response categories; see section Middle Categories in Dominance Items).

Furthermore, this article highlights the flexibility of the proposed general IRTree framework, which can be considered a modular system with three independent components that can be combined as desired. One component of IRTree models is the psychological theory concerning the decomposition of ordinal rating responses into sub-decisions. By specifying the number and structure of the sub-decisions, theory-driven hypotheses on the logical sequence of cognitive processing stages can be defined (see Figure 1). A second component is the definition of the response processes that contribute to the individual sub-decisions. Since various processes can be assigned to the pseudo-items separately from each other, personal characteristics may be involved in one or more pseudo-items, and pseudo-items may depend on one or more processes. Using the new DI-MIRT model, it is now possible to define a third IRTree component independently of the other two, which is the choice of process-specific IRFs. The individual response processes can be parameterized by dominance or ideal point models, and they can be freely combined both within and between pseudo-items.

In the remainder of this article, we will illustrate this modular system and present exemplary IRTree models that differ in terms of the selection and combination of the three components. To this end, we firstly derive the new DI-MIRT model of co-occurring processes from existing models of the divide-by-total framework. Secondly, we introduce IRTree models in which dominance and ideal point processes co-occur within sub-decisions. Thirdly, a simulation study on parameter recovery and model selection is presented. Then, the utility of the new approach is illustrated by two empirical examples, the first one focusing on the investigation of the relative importance of co-occurring processes in IRTree sub-decisions, and the second one using response time data

to provide a construct validation of the parameter estimates. We conclude with a discussion of the results.

### Existing Dominance and Ideal Point Divide-By-Total Models

The divide-by-total framework contains both dominance and ideal point models. For both of them holds that the category probabilities of ordinal responses  $Y \in \{0, \dots, K\}$  are defined as the ratio of category-specific components divided by the sum of the  $K + 1$  components of all available categories, so that the probabilities across categories sum up to 1. For modeling item responses of person  $v = 1, \dots, N$  to item  $i = 1, \dots, I$  under the dominance assumption (D), divide-by-total models can take the form

$$p^{(D)}(Y_{vi} = y_{vi}) = \frac{\omega_{viy}}{\sum_{j=0}^K \omega_{vij}} = \frac{\exp(\eta_{viy})}{\sum_{j=0}^K \exp(\eta_{vij})}, \quad (2)$$

where  $\eta_{viy}$  is a linear combination of person and item parameters.

A prominent member of dominance divide-by-total models is the generalized partial credit model (GPCM; Muraki, 1992), which is given by

$$p^{(D)}(Y_{vi} = y_{vi} \mid \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\exp \left[ \alpha_i \left( s_y \theta_v - \sum_{k=0}^y \beta_{ik} \right) \right]}{\sum_{j=0}^K \exp \left[ \alpha_i \left( s_j \theta_v - \sum_{k=0}^j \beta_{ik} \right) \right]}, \quad (3)$$

with  $\beta_{i0} := 0$  and  $s_y = y$ .  $\theta_v$  denotes the person-specific trait level,  $\alpha_i$  the item-specific discrimination parameter, and  $\beta_{ik}$  the item- and category-specific thresholds (or difficulties). The threshold parameters can be rewritten as  $\beta_{ik} = \beta_i + \zeta_{ik}$ , where  $\beta_i$  denotes the item location and is defined as  $\sum_{k=1}^K \beta_{ik}/K$  and  $\zeta_{ik}$  denotes the category-specific deviations. If the thresholds  $\beta_{ik}$  are ordered across the ordinal categories, each category has a section on the latent trait continuum for which the probability to be chosen is higher than the probabilities of all other categories. The scoring weights  $\mathbf{s}$  define the relation between trait and response categories and are fixed to  $s_y = y$  in the GPCM, reflecting that the response categories are ordered and that higher trait levels are associated with higher categories. However, they can be set to any

other values depending on the theoretical assumptions, or can be estimated like in the nominal response model for categorical responses (Bock, 1972; Thissen et al., 2010). Note that the choice of scoring weights depends on the assumption of how the trait influences the selection of categories, but does not change the underlying dominance assumption. Figure 3A shows exemplary category probability curves for an item on a four-point scale under the GPCM.

For modeling item responses under the ideal point assumption (I), divide-by-total models can take the form

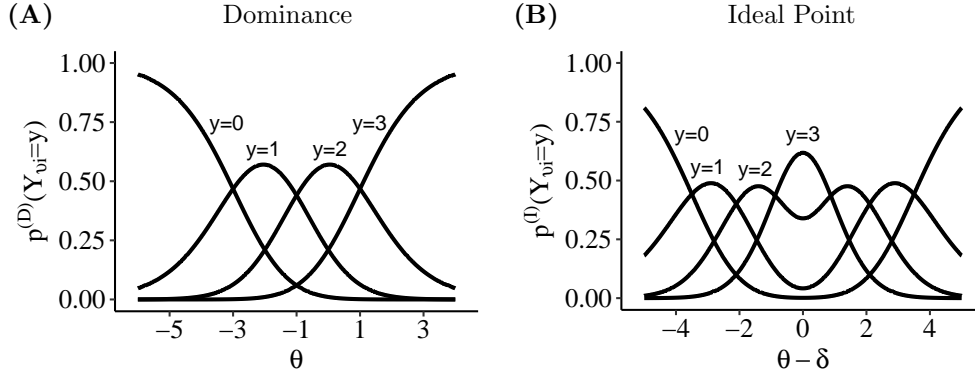
$$p^{(I)}(Y_{vi} = y_{vi}) = \frac{\omega_{viy}}{\sum_{j=0}^K \omega_{vij}} = \frac{\exp(\eta_{1viy}) + \exp(\eta_{2viy})}{\sum_{j=0}^K (\exp(\eta_{1vij}) + \exp(\eta_{2vij}))}, \quad (4)$$

where  $\eta_{1viy}$  and  $\eta_{2viy}$  are linear combinations of person and item parameters. The category-specific components  $\omega_{viy}$  are defined as the sum of two exponential terms since it is assumed that each response category of a rating scale is composed of two unobservable, latent categories. Each two associated latent categories reflect the perspectives from above and from below the item location, respectively. For instance, the observable categories "agree" and "disagree" of a dichotomous item correspond to the latent categories "disagree from below", "agree from below", "agree from above", and "disagree from above". Thus, ideal point divide-by-total models account for two different reasons for which respondents can select a specific category, such as disagreement being chosen because of having a much higher or a much lower ideal point compared to the item location. By adding up the probabilities of selecting a category from below and from above, probabilities of the observable categories are obtained.

The generalized graded unfolding model (GGUM; Roberts et al., 2000) belongs to the class of ideal point divide-by-total models and is given by

$$\begin{aligned}
p^{(I)}(Y_{vi} = y_{vi} \mid \mathbf{s}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \boldsymbol{\xi}) = & \\
& \frac{\exp \left[ \lambda_i \left( s_y (\theta_v - \delta_i) - \sum_{k=0}^y \xi_{ik} \right) \right] + \exp \left[ \lambda_i \left( (M - s_y) (\theta_v - \delta_i) - \sum_{k=0}^y \xi_{ik} \right) \right]}{\sum_{j=0}^K \left\{ \exp \left[ \lambda_i \left( s_j (\theta_v - \delta_i) - \sum_{k=0}^j \xi_{ik} \right) \right] + \exp \left[ \lambda_i \left( (M - s_j) (\theta_v - \delta_i) - \sum_{k=0}^j \xi_{ik} \right) \right] \right\}} \quad (5)
\end{aligned}$$

with  $\xi_{i0} := 0$ ,  $M = 2K + 1$  and  $s_y = y$ .  $\theta_v$  denotes the person's trait level (ideal point),  $\delta_i$  the item location,  $\lambda_i$  the discrimination parameter, and  $\xi_{ik}$  the category-specific threshold. If the thresholds  $\xi_{ik}$  are  $< 0$  and ordered across the ordinal categories, all observable categories have sections on the latent trait continuum for which the probability to be chosen is higher than for the other categories. Note that the parameterization of each exponential term has high similarity to the GPCM, which in fact causes the category probability curves of the  $M + 1$  latent categories to take the form of a GPCM. The first exponential term in the numerator and denominator of the GGUM corresponds to latent response categories from below, whereas the second term corresponds to latent categories from above. Each two associated latent categories differ only in their scoring weights of the person-item proximity ( $\theta_v - \delta_i$ ), which are defined as  $y$  and  $(M - y)$ , respectively. Thus, the scoring weights of latent categories from below increase across categories  $(0, \dots, K)$ , whereas they decrease to the same extent for latent categories from above  $(M, \dots, K + 1)$ . The observable category probabilities are symmetrical about the point  $(\theta_v - \delta_i) = 0$ , which implies that selecting response category  $y$  is equally likely for a positive or negative deviation of a respondent's trait level from the item location. As is the case for dominance models, the scoring weights of ideal point divide-by-total models can be fixed to any values, which would reflect different hypotheses on the relation between trait and categories, without changing the ideal point rationale. Figure 3B shows exemplary category probability curves for an item on a four-point scale under the GGUM.

**Figure 3***Category Probability Curves for Responding to Four-Point Rating Items*

*Note.* Category probabilities under the dominance assumption of the GPCM (A) and under the ideal point assumption of the GGUM (B). Discrimination parameters  $\alpha$  and  $\lambda$  are set to 1; the thresholds are set as follows:  $\beta_1 = -3$ ;  $\beta_2 = -1$ ;  $\beta_3 = 1$ ;  $\xi_1 = -3.5$ ;  $\xi_2 = -2.1$ ;  $\xi_3 = -0.6$ .

### Co-Occurring Dominance and Ideal Point Processes

The novel DI-MIRT model of co-occurring processes is a multi-process generalization of dominance and ideal point divide-by-total models and includes both of them as special cases. In order to combine response processes described by those two models, the definitions of dominance and ideal point approaches must be brought into the same format. As described above, ideal point divide-by-total models consist of the sum of two exponential terms, which correspond to the two underlying latent categories together defining the probability distribution of observable categories. For dominance models, in contrast, the probability distribution of observable categories can be modeled directly. Nonetheless, such models can likewise be displayed in the form of two added components, by applying the single linear parameter combination of a dominance model (which is  $\eta_{vij}$  in Equation 2) to both exponential terms in the ideal point formulation ( $\eta_{1vij}$  and  $\eta_{2vij}$  in Equation 4). Consequently, all category-specific components  $\omega_{vij}$  are simply doubled both in the numerator and denominator, which does not affect the probability distribution across categories, so an equivalent model results. Metaphorically speaking, each observable response category is artificially divided into two latent categories, which are selected with equal

probability. Thereby, ideal point and dominance models can be expressed in the same form and are represented by two added components. If several response processes  $r \in \{1, \dots, R\}$ , no matter if dominance or ideal point processes, are to be aggregated to a common probability distribution, the respective linear parameter combinations can simply be added within each of the two exponential terms. The resulting DI-MIRT model is given by

$$p(Y_{vi} = y_{vi}) = \frac{\exp\left(\sum_{r=1}^R \eta_{1viyr}\right) + \exp\left(\sum_{r=1}^R \eta_{2viyr}\right)}{\sum_{j=0}^K \left[ \exp\left(\sum_{r=1}^R \eta_{1vijr}\right) + \exp\left(\sum_{r=1}^R \eta_{2vijr}\right) \right]}. \quad (6)$$

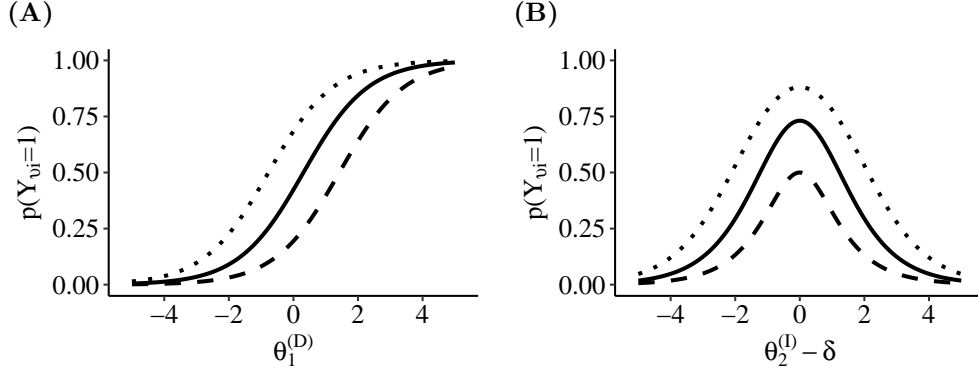
With this general formulation, the co-occurrence of several dominance processes, several ideal point processes, or a combination of both can be modeled in a consistent way. Note that the two linear parameter combinations do not differ for dominance response processes ( $\eta_{1viyr} = \eta_{2viyr}$ ), whereas the parameterizations for ideal point processes differ in their scoring weights (see Equation 5). In the further course of the article, we use the GPCM for modeling dominance processes and the GGUM for ideal point processes. However, the individual processes can be defined by any unidimensional dominance or ideal point IRT model which can be represented in the form of divide-by-total models (as defined in Equation 2 and Equation 4). Figure 4 illustrates the co-occurrence of one dominance and one ideal point process for an exemplary binary item. The higher a person's dominance trait level ( $\theta_1^{(D)}$ ) and the higher the proximity of a person's ideal point to the item location ( $|\theta_2^{(I)} - \delta|$ ), the higher the probability of endorsing the item.

## Identification

The DI-MIRT model is a suitable theoretical model of how co-occurring processes jointly determine item responding, though it is highly parameterized and not identified without certain constraints, whenever two or more processes are to be considered. Given that the linear parameter combinations of the co-occurring processes ( $\eta_{1viyr}$  and  $\eta_{2viyr}$ ) each consist of a person-specific trait and item-specific category thresholds, some restriction may arise with respect to estimating those two kinds of parameters.

**Figure 4**

*Probability Curves for Endorsing a Binary Item Under the DI-MIRT Model*



*Note.* One response process follows the dominance assumption parameterized by the GPCM ( $\theta_1^{(D)}$ ) and the other process follows the ideal point assumption parameterized by the GGUM ( $\theta_2^{(I)}$ ). A: Probability curves for fixed  $|\theta_2^{(I)} - \delta|$  of .5 (dotted line), 1.5 (solid line), and 2.5 (dashed line). B: Probability curves for fixed  $\theta_1^{(D)}$  of 1 (dotted line), 0 (solid line), and -1 (dashed line). The other parameters are set as follows:  $\alpha = 1$ ,  $\lambda = 1$ ,  $\beta_1 = 0$ ,  $\xi_1 = -1$ ,  $\mathbf{s}^{(D)} = (0, 1)$ , and  $\mathbf{s}^{(I)} = (0, 1)$ .

Firstly, the threshold parameters of several processes cannot be separated, so only one common threshold per category can be estimated. This common threshold, which we call *category intercept*, is a weighted linear combination of the threshold parameters of the co-occurring processes. It is, therefore, not possible to examine the individual contributions of the involved processes to the size of each category intercept, that is, to why specific response categories are selected more or less frequently. For instance, for dominance processes modeled by the GPCM and ideal point processes modeled by the GGUM, the linear parameter combinations are defined as

$$\sum_{r=1}^R \eta_{1viyr} = \sum_{r=1}^R \left[ (\alpha_{ir} s_{yr} \theta_{vr})^{m_r} + (\lambda_{ir} s_{yr} (\theta_{vr} - \delta_{ir}))^{(1-m_r)} \right] - \sum_{k=0}^y \tau_{ik} \quad (7)$$

and

$$\sum_{r=1}^R \eta_{2viyr} = \sum_{r=1}^R \left[ (\alpha_{ir} s_{yr} \theta_{vr})^{m_r} + (\lambda_{ir} (M - s_{yr}) (\theta_{vr} - \delta_{ir}))^{(1-m_r)} \right] - \sum_{k=0}^y \tau_{ik}, \quad (8)$$

where  $m_r = 1$  if process  $r$  is a dominance process and  $m_r = 0$  if it is an ideal point process.  $\tau_{ik}$  is the category intercept of the  $R$  processes and is given by

$$\tau_{ik} = \sum_{r=1}^R (\alpha_{ir} \beta_{ikr})^{m_r} + (\lambda_{ir} \xi_{ikr})^{(1-m_r)}. \quad (9)$$

Note that this constraint of only the common category intercept of several processes being estimated applies to existing MIRT models as well (such as the multidimensional nominal response model; Bolt & Johnson, 2009), though it is implicitly captured in the model formulations by defining only one threshold in the first place.

The second constraint of the DI-MIRT model is that the respondents' trait levels can only be separated from each other if the scoring weights differ in at least one of the two exponential terms. Therefore, two (or more) dominance processes or two (or more) ideal point processes must be defined to affect the response categories in different ways. Again, this also applies to other MIRT models. For instance, in multidimensional models for the analysis of ERS, the ERS-based process typically gets assigned the scoring weights  $(1, 0, \dots, 0, 1)$ , reflecting the assumption that only the outermost two categories are affected, and thus can be separated from the ordinal influence of a trait with scoring weights  $(0, \dots, K)$ .

In addition to the above remarks, it should be noted that the scale of the latent continuum is not per se identified in the DI-MIRT model – as is the case for any other IRT model. When estimating the DI-MIRT model, the location, the variability, and the orientation of the continuum have to be fixed. The identification of the location is required in all IRT models and is typically done by setting the mean of the latent trait distribution to zero. Models with discrimination parameters, such as the GPCM, additionally require fixing the variability, which is often done by setting the variance of the trait distribution to one. In ideal point models, such as the GGUM, the orientation (or sign) of the continuum is unknown and therefore has to be fixed. The non-identified orientation is due to the fact that the estimation of the trait



levels and item locations in ideal point models is based on their proximity, that is, the pairwise distances on the common latent continuum. Thus, two sets of parameters result in the same likelihood, whereby the two parameter solutions only differ in the signs of the person-specific trait levels and item-specific locations. Importantly, both solutions are equally correct; only the meaning of the latent continuum changes, so it is up to the researcher to decide which of the two parameter sets corresponds to the interpretation of the latent continuum that is more intuitive. The practical specification of the continuum orientation is described below in the context of the simulation study (see Estimation and Analysis).

### Probability-Based Formulation

The DI-MIRT formulation as divide-by-total model given by Equation 6 can be rewritten as a model which aggregates processes at the level of process-specific category probabilities. Let  $\mathbf{p}^{(r)}(Y_{vi} = y_{vi})$  denote the vector of probabilities for responding to each of the  $K + 1$  categories of an item for response process  $r$ . The joint probability for  $R$  co-occurring processes can then be obtained by passing the process-specific probabilities to a probability-aggregating function  $\sigma$ , which is defined as

$$p(Y_{vi} = y_{vi}) = \sigma_y \left[ \mathbf{p}^{(1)}(Y_{vi} = y_{vi}), \dots, \mathbf{p}^{(R)}(Y_{vi} = y_{vi}) \right] = \text{softmax} \left[ \sum_{r=1}^R \log \left( \mathbf{p}^{(r)}(Y_{vi} = y_{vi}) \right) \right]_{(y+1)}, \quad (10)$$

with softmax being a normalized exponential function transforming a  $Z$ -dimensional vector  $\mathbf{x}$  into a vector of probabilities summing up to 1. Position  $z$  of this probability vector is given by

$$\text{softmax}[\mathbf{x}]_z = \frac{\exp(x_z)}{\sum_{w=1}^Z \exp(x_w)} \quad \text{for } z = 1, \dots, Z. \quad (11)$$

This formulation of the DI-MIRT model with aggregation of response processes on the level of category probabilities is equivalent to the aggregation on the level of linear parameter combinations derived above. Thus, even though these two formulations seem to differ regarding their

theoretical assumptions (the aggregation at the level of linear parameter combinations suggests that multiple response processes are simultaneously active; the probability-based approach suggests separate cognitive processing and subsequent aggregation), they cannot be distinguished statistically.

Note, however, that the probability-based model is more general and not restricted to divide-by-total models, since the category-specific probabilities of a process could result from any kind of IRT model (e.g., difference models like the graded response model; Samejima, 1969). Nonetheless, depending on the choice of process-specific models, the estimated parameters might have unintuitive interpretations (e.g., the threshold parameters in a co-occurring model including one process defined by a difference model and another process defined by a divide-by-total model) and different constraints might be necessary in order to identify the models. In this article, we will therefore only refer to processes defined by divide-by-total models, for which both formulations are equivalent.

### **IRTree Models of Co-Occurring Processes**

IRTree models allow to separate the influences of multiple latent traits, which can be involved in the response selection both sequentially across sub-decisions and as co-occurring processes within sub-decisions. The modular system of IRTree models offers high flexibility for the specification of theoretical assumptions with respect to three components: (a) the decomposition of ordinal response into sub-decisions, (b) the response processes involved in such sub-decisions, and (c) the IRFs of the individual processes of each pseudo-item. In this article, we introduce two exemplary IRTree models from the field of RS modeling in which these three components are combined in different ways, and in which dominance and ideal point processes co-occur in at least one pseudo-item. The first model addresses influences of RS on responding to ideal point items; the second one addresses the selection of middle response categories in dominance items. The two models are formally outlined in the following two sections (Response Style Analysis in Ideal Point Items and Middle Categories in Dominance Items, respectively) and applied to empirical data sets in the section Empirical Applications. Even though these exemplary models cover only a part of all conceivable IRTree configurations that arise from the DI-MIRT frame-

work, they illustrate the advantages of the new parameterization for IRTree pseudo-items. Our proposed approach is a rather general one and can be easily adapted to differently structured trees and response processes outside RS modeling. Further, we will only refer to pseudo-items with a maximum of two response processes, although the new approach would allow modeling any number of co-occurring response processes. Within IRTree models, however, we consider this as a reasonable restriction since different processes, like MRS and ERS, are assumed to be sequentially involved in item responding and do not affect the same sub-decisions.

### Response Style Analysis in Ideal Point Items

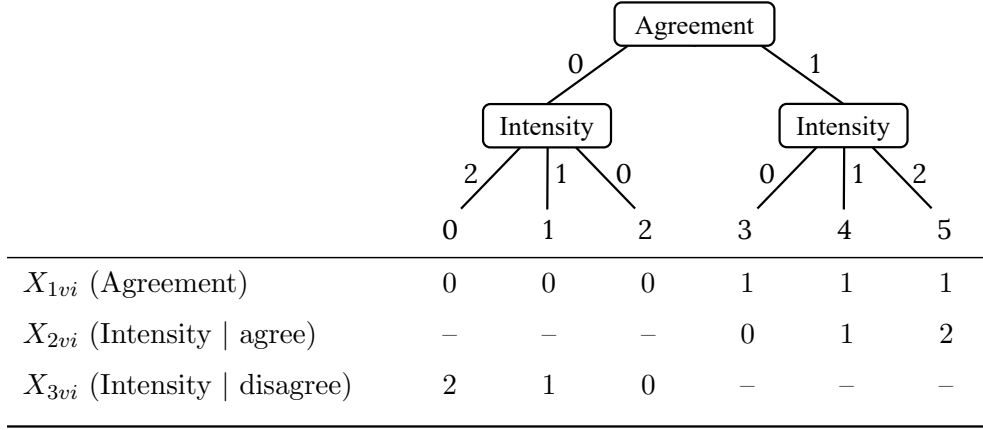
The first application concerns IRTree models for the analysis of ideal point items while controlling for RS. Assuming that some sub-decisions within the response selection depend on both the trait (i.e., an ideal point process) and a RS (i.e., a dominance process), the model of co-occurring processes is necessary for modeling the respective pseudo-items. We illustrate the analysis of this kind of item response data by an IRTree model of six-point rating scale items, which are decomposed into two sub-decisions of agreement and intensity, as depicted in Figure 5. The probability of an ordinal response  $Y_{vi} \in \{0, \dots, 5\}$  of person  $v = 1, \dots, N$  to item  $i = 1, \dots, I$  is the product of the probabilities of responses to the pseudo-items  $X_{hvi}$ , where one pseudo-item reflects an agreement decision ( $h = 1$ ) and two pseudo-items reflect the decisions regarding the intensity of responses conditional on the agreement judgment ( $h = 2$  and  $h = 3$ ):

$$p(Y_{vi} = y_{vi}) = p(X_{1vi} = x_{1vi}) \times p(X_{2vi} = x_{2vi})^{x_{1vi}} \times p(X_{3vi} = x_{3vi})^{(1-x_{1vi})}. \quad (12)$$

Under the assumption that all sub-decisions are dependent on the substantive trait  $\theta$ , and that the intensity judgments are additionally affected by the ERS  $\eta$ , the probabilities of the three pseudo-items can be specified as follows:

**Figure 5**

*Tree Diagram and Definition of Pseudo-Items for Responses to Six-Point Rating Items*



*Note.* Pseudo-items that are missing by design are marked with '–'.

$$\begin{aligned}
 p(X_{1vi} = x_{1vi}) &= p^{(I)}(x_{1vi} | \mathbf{s} = (0, 1), \theta_v, \lambda_{1i}, \delta_i, \xi_{1i}) \\
 p(X_{2vi} = x_{2vi}) &= \sigma_x \left[ p^{(D)}(x_{2vi} | \mathbf{s} = (0, 1, 2), \eta_v, \alpha_i, \beta_{1i}), p^{(I)}(x_{2vi} | \mathbf{s} = (0, 1, 2), \theta_v, \lambda_{2i}, \delta_i, \xi_{2i}) \right] \\
 p(X_{3vi} = x_{3vi}) &= \sigma_x \left[ p^{(D)}(x_{3vi} | \mathbf{s} = (0, 1, 2), \eta_v, \alpha_i, \beta_{2i}), p^{(I)}(x_{3vi} | \mathbf{s} = (2, 1, 0), \theta_v, \lambda_{2i}, \delta_i, \xi_{3i}) \right],
 \end{aligned} \tag{13}$$

where  $p^{(D)}$  denotes response probabilities under the GPCM as given in Equation 3,  $p^{(I)}$  denotes probabilities under the GGUM as given in Equation 5, and  $\sigma$  denotes the probability-aggregating function given in Equation 10. As the sub-decision of agreement depends on a trait-based ideal point process and the intensity decision comprises co-occurring trait and ERS processes, this IRTree structure is abbreviated as  $I_{\theta-D_{\eta}I_{\theta}}$  in the following.

The scoring weights  $\mathbf{s}$  of the trait-based ideal point process differ between the two intensity pseudo-items to account for the fact that high proximity of a respondent's ideal point and the item's location (i.e., a small difference of  $\theta_v$  and  $\delta_i$ ) increases the probability of intense agreement but reduces the probability of intense disagreement. As such intensity scoring weights relate to the definition of the pseudo-items in Figure 5 from inner to outer ordinal categories of the scale, the respective first weights refer to the least intense ordinal categories (3 and 2 for

agreement and disagreement, respectively), followed by the weights for moderately intense (4 and 1) and the most intense (5 and 0) categories. Further, the ordinal definition of the ERS-based dominance process implies that a preference for the extreme categories and a preference for the midscale categories are opposite poles of a common trait. Thus, positive ERS levels increase the probability to select extreme categories and decrease the probability of midscale categories, while it is the other way around for negative ERS levels. Also note that the thresholds of the co-occurring processes within the intensity pseudo-items cannot be separated, meaning that only one category intercept can be estimated. Such category intercepts  $\tau$  are defined as given in Equation 9: For pseudo-item  $X_{1vi}$ , only one response process is defined, so that the intercept  $\tau_{1i}$  is simply given by  $\lambda_{1i}\xi_{1i}$ . For pseudo-item  $X_{2vi}$ , the intercept is  $\tau_{2i} = \alpha_i\beta_{1i} + \lambda_{2i}\xi_{2i}$ , and for pseudo-item  $X_{3vi}$ , it is  $\tau_{3i} = \alpha_i\beta_{2i} + \lambda_{2i}\xi_{3i}$ .

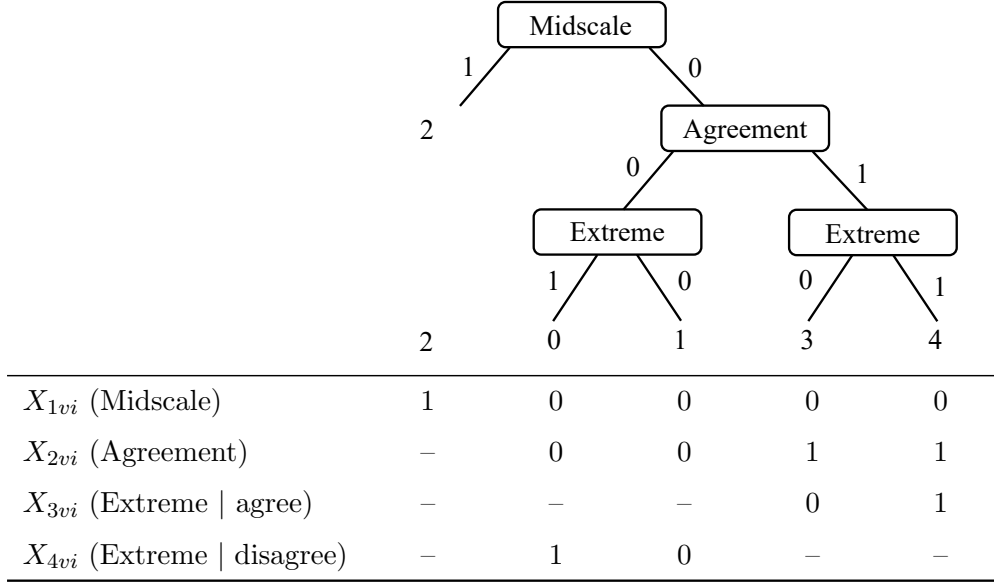
### Middle Categories in Dominance Items

The second application relates to IRTree sub-decisions of midscale versus non-midscale responding in dominance items. There is an ongoing discussion in the literature on whether middle categories are used as part of the ordinal scale and reflect a neutral attitude of respondents, or whether they are rather considered as a non-response option and selected to avoid providing personal information (e.g., Kalton et al., 1980; Nowlis et al., 2002; Sturgis et al., 2014; Tijnstra & Bolsinova, in press; Tijnstra et al., 2018). The first interpretation implies a trait-based response selection; the latter one indicates that such decisions are based on another trait, which could be referred to as MRS. Thus, it seems reasonable to consider both kinds of response processes in a co-occurring model in order to examine their relative importance for the selection of middle categories.

For such sub-decisions of midscale responding, the substantive trait behaves like an ideal point process, despite the fact the item generally follows the dominance rationale. Therefore, trait-based agreement is modeled as a dominance process, whereas trait-based midscale responding as an ideal point process. This unintuitive property is due to the fact that only respondents who have moderately high substantive trait levels in relation to the item location are assumed to select middle categories as an expression of a neutral opinion. In contrast, respondents having

**Figure 6**

*Tree Diagram and Definition of Pseudo-Items for Responses to Five-Point Rating Items.*



*Note.* Pseudo-items that are missing by design are marked with '–'.

a very high or very low trait level are unlikely to select neutral response categories, because they are assumed to have a clear-cut opinion regarding the item content. Accordingly, if such a trait-based ideal point process is to be modeled to co-occur with a MRS-based dominance process, the DI-MIRT model is required.

This use case of the DI-MIRT model is illustrated by an IRTree model of items on a five-point rating scale with the three sub-decisions of midscale responding, agreement, and extreme responding, as depicted in Figure 6. The probability of an ordinal response  $Y_{vi} \in \{0, \dots, 4\}$  is the product of the conditional pseudo-item probabilities  $X_{hvi}$  and is given by:

$$p(Y_{vi} = y_{vi}) = p(X_{1vi} = x_{1vi}) \times \left[ p(X_{2vi} = x_{2vi}) \times p(X_{3vi} = x_{3vi})^{x_{2vi}} \times p(X_{4vi} = x_{4vi})^{(1-x_{2vi})} \right]^{(1-x_{1vi})}, \quad (14)$$

Assuming that the decision of midscale responding depends on the substantive trait  $\theta$  and the MRS  $\eta_1$ , that agreement is solely trait-based, and that extreme responding depends on the

trait and the ERS  $\eta_2$ , the pseudo-item probabilities can be defined as follows:

$$\begin{aligned}
p(X_{1vi} = x_{1vi}) &= \sigma_x \left[ p^{(D)}(x_{1vi} | \mathbf{s} = (0, 1), \eta_{1v}, \alpha_{1i}, \beta_{1i}), p^{(I)}(x_{1vi} | \mathbf{s} = (0, 1), \theta_v, \lambda_i, \delta_i = \beta_{2i}, \xi_i) \right] \\
p(X_{2vi} = x_{2vi}) &= p^{(D)}(x_{2vi} | \mathbf{s} = (0, 1), \theta_v, \alpha_{2i}, \beta_{2i}) \\
p(X_{3vi} = x_{3vi}) &= \sigma_x \left[ p^{(D)}(x_{3vi} | \mathbf{s} = (0, 1), \eta_{2v}, \alpha_{3i}, \beta_{3i}), p^{(D)}(x_{3vi} | \mathbf{s} = (0, 1), \theta_v, \alpha_{4i}, \beta_{4i}) \right] \\
p(X_{4vi} = x_{4vi}) &= \sigma_x \left[ p^{(D)}(x_{4vi} | \mathbf{s} = (0, 1), \eta_{2v}, \alpha_{3i}, \beta_{5i}), p^{(D)}(x_{4vi} | \mathbf{s} = (1, 0), \theta_v, \alpha_{4i}, \beta_{6i}) \right],
\end{aligned} \tag{15}$$

where  $p^{(D)}$  denotes the GPCM,  $p^{(I)}$  the GGUM, and  $\sigma$  the probability-aggregating function. Note that the DI-MIRT parameterization is used in two ways, once for modeling the co-occurrence of a dominance and an ideal point process in the midscale pseudo-item, and once to model two dominance processes in the extreme pseudo-items. This IRTree structure is abbreviated  $D_{\eta_1} I_{\theta} - D_{\theta} - D_{\eta_2} D_{\theta}$ .

As before, the thresholds of co-occurring response processes within one pseudo-item cannot be separated so that common category intercepts are defined for all pseudo-items. For pseudo-item  $X_{1vi}$ , the intercept is  $\tau_{1i} = \alpha_{1i}\beta_{1i} + \lambda_i\xi_i$ , for pseudo-item  $X_{2vi}$  it is  $\tau_{2i} = \alpha_{2i}\beta_{2i}$ , for pseudo-item  $X_{3vi}$  it is  $\tau_{3i} = \alpha_{3i}\beta_{3i} + \alpha_{4i}\beta_{4i}$ , and for pseudo-item  $X_{4vi}$  it is  $\tau_{4i} = \alpha_{3i}\beta_{5i} + \alpha_{4i}\beta_{6i}$ . Importantly, the item location of the trait-based process of midscale responding  $\delta_i$  is not estimated independently, but set equal to the threshold of agreement  $\beta_{2i}$ . This equality constraint implies that respondents whose trait levels are equal to the agreement difficulty parameter (a) have maximal ambiguity regarding the decision to agree or disagree (see pseudo-item  $X_{2vi}$ ), and (b) have maximal probability for a trait-based selection of the middle category (see pseudo-item  $X_{1vi}$ ). The larger the distance of the respondents' trait levels to the item difficulty, the more clear-cut the agreement decisions and the less likely midscale responses are.

### Simulation Study

A simulation study was conducted to evaluate the parameter recovery and model fit of IRTree pseudo-items of co-occurring processes modeled by the DI-MIRT model. The study was based on

the  $I_\theta$ - $D_\eta$ - $I_\theta$  IRTree model described in the section Response Style Analysis in Ideal Point Items, which analyzes ideal point items on a six-point rating scale while incorporating an ERS influence in the intensity sub-decisions. We choose this model for the simulations since all six response categories were influenced by co-occurring dominance and ideal point processes. This model of co-occurring response processes was compared to two models of sequential processes, in which the intensity pseudo-items were unidimensional and dependent on only one of the two processes, that is, either ERS-based (trait-ERS model of sequential processes;  $I_\theta$ - $D_\eta$ ) or trait-based (trait-trait model of sequential processes;  $I_\theta$ - $I_\theta$ ). We evaluated how the co-occurring model performed when it was the true data-generating model, and when it was over-parameterized and fitted to data generated under the models of sequential processes, which are both nested within it. Further, we evaluated how the models of sequential processes performed when fitted to data generated by the co-occurring model, meaning that one of the two intensity processes was ignored in the analysis.

### Data Generation

Item response data were generated for each of the three models, which are described by Equation 12 and Equation 13 ( $I_\theta$ - $D_\eta$ - $I_\theta$  model of co-occurring processes), or by special cases with unidimensional pseudo-items ( $I_\theta$ - $D_\eta$  trait-ERS model and  $I_\theta$ - $I_\theta$  trait-trait model of sequential processes). 100 replications were conducted for each of two sample sizes  $N$ , set to 500 and 1000, and two questionnaire lengths  $I$ , set to 10 and 20. The person-specific trait levels  $\theta_v$  and ERS levels  $\eta_v$  were sampled from independent standard normal distributions. The discrimination parameters  $\alpha_i$ ,  $\lambda_{1i}$ , and  $\lambda_{2i}$  were drawn from  $LogN(0, 0.25)$ . The item locations  $\delta_i$  were drawn from a uniform distribution  $U(-3, 3)$ . The thresholds of ideal point processes were sampled from the distributions  $N(-2.2, 0.2)$ ,  $N(-1.3, 0.2)$ ,  $N(-1, 0.2)$ ,  $N(-0.8, 0.2)$ , and  $N(-0.2, 0.2)$ . The means of these threshold distributions were ordered across ordinal response categories, so that the first two correspond to the thresholds of intense disagreement ( $\xi_{3i}$ ), the third to the threshold of the agreement pseudo-item ( $\xi_{1i}$ ), and the last two to the two thresholds of intense agreement ( $\xi_{2i}$ ). For dominance processes, the thresholds  $\beta_{1i}$  and  $\beta_{2i}$  were defined as the sum of item-specific locations  $\beta_i$ , which were generated from  $U(-1, 1)$ , and the category-specific de-



viations  $\zeta_{ik}$ , which were drawn from  $N(-0.5, 0.2)$  and  $N(0.5, 0.2)$ . Item responses were sampled according to the model-implied probabilities.<sup>3</sup>

### Estimation and Analysis

All analyses were conducted in R (R Core Team, 2023). Bayesian parameter estimation was performed since the proposed IRTree models with DI-MIRT parameterization are comparably complex and their estimation would probably not be possible within the frequentist framework. We used the software program Stan (Stan Development Team, 2023) and the R package CmdStanR (Gabry et al., 2023). For each generated data set, the three models ( $I_{\theta-D_{\eta}I_{\theta}}$ ,  $I_{\theta-D_{\eta}}$ , and  $I_{\theta-I_{\theta}}$ ) were fitted. Priors were set as follows:  $\alpha \sim \text{Gamma}(1.5, 1.5)$ ,  $\lambda \sim \text{Gamma}(1.5, 1.5)$ , and  $\tau \sim N(0, 5)$ . The parameters  $\delta$  were given a hierarchical prior with a  $N(0, 5)$  hyperprior for the mean and nonnegative  $N(0, 5)$  for the standard deviation. The distributions of  $\theta$  and  $\eta$  were set to  $N(0, 1)$  for identifying the models.

Furthermore, we set initial values for the Markov chain Monte Carlo (MCMC) chains in all models. Firstly, this was important to avoid MCMC chains getting stuck in local maxima. The same applies to the GGUM, for which Roberts et al. (2000) proposed to fit a constrained model first, and to use the estimates of this model as initial values for the full model. We followed this procedure and defined three constrained models (corresponding to the three full models), in which the discrimination parameters ( $\alpha$  and  $\lambda$ ) and category intercepts ( $\tau$ ) were set equal across items. Secondly, setting initial values also allowed to specify the orientation of the latent continuum, which otherwise would not be identified. To this end, the initial values of the item locations were set in accordance with one of the two possible scale orientations. Thereby, the chains only explore that part of the posterior distribution that aligns with this parameter solution and do not jump to the alternative parameter set. As suggested by Liu and Wang (2016), the signs of the item locations were treated as known, which is why such parameters were initialized with values 1 or -1, depending on the signs of the respective generated data set (for empirical data, the signs of the item locations are obviously not known, so content

---

<sup>3</sup>The R code for generating data sets can be found in the OSF project; [https://osf.io/yu4gx/?view\\_only=50fc21d10d52414aaeece310d680fc0e](https://osf.io/yu4gx/?view_only=50fc21d10d52414aaeece310d680fc0e).

knowledge can inform the selection of one of the two possible scale orientations; a slightly different procedure for setting the initial values is then used as described in the Empirical Applications). For the constrained models, one chain with 500 warmup iterations and 500 post-warmup iterations was run to derive approximate estimates for the model parameters. The expected a posteriori (EAP) estimates were then used to create initial values for fitting the full model.

For the full model, four chains with 500 warmup iterations and 1000 post-warmup iterations were run. To ensure model convergence and enough independent posterior samples for the estimation of each parameter, the Gelman-Rubin statistic  $\hat{R}$  and the effective sample size were evaluated (for more information on these diagnostics, see Vehtari et al., 2021). If at least one model parameter had an  $\hat{R}$  value greater than 1.05 or either the bulk or tail effective sample size was smaller than 100, more samples were drawn (in steps of 500, up to 3000 post-warmup iterations). By this procedure, accurate estimates were achieved for all models while keeping the computation time reasonable for models which provided good diagnostic values after fewer samples.<sup>4</sup>

It is important to note that despite the careful choice of initial values and the interim step of fitting a constrained model, by chance, some MCMC chains may move to either a local maximum or to the area of the posterior distribution which corresponds to the solution with inverted scale orientation. This becomes apparent in that the model does not converge or that the signs of estimated item locations are inverted compared to the initial values. Since this only occurs in very few instances, the model can simply be re-fitted with a different seed. In the simulation study, we ensured that all models converged to the solution of the scale orientation corresponding to that of the data generation to ensure sensible results for the parameter recovery.

The fitted models (i.e., the co-occurring model and the two models of sequential processes) were compared regarding their parameter recovery by mean absolute bias (MAB) of the EAP point estimates. Further, out-of-sample model fit was compared by an approximation of leave-one-out cross-validation based on Pareto smoothed importance sampling (LOO; Vehtari et al.,

---

<sup>4</sup>The Stan model code and an R script illustrating the estimation procedure can be found in the OSF project. The R script also shows how traceplots can be used as an additional check of model convergence.

**Table 1***Recovery of Person Parameters by MAB*

Generation	Analysis	Trait $\theta$		ERS $\eta$	
		<i>I</i> 10	<i>I</i> 20	<i>I</i> 10	<i>I</i> 20
$I_{\theta-D_{\eta}I_{\theta}}$	$I_{\theta-D_{\eta}I_{\theta}}$	0.346	0.249	0.395	0.296
	$I_{\theta-D_{\eta}}$	0.496	0.375	0.434	0.352
	$I_{\theta-I_{\theta}}$	0.377	0.286		
$I_{\theta-D_{\eta}}$	$I_{\theta-D_{\eta}}$	0.497	0.377	0.351	0.264
	$I_{\theta-D_{\eta}I_{\theta}}$	0.502	0.379	0.352	0.265
$I_{\theta-I_{\theta}}$	$I_{\theta-I_{\theta}}$	0.309	0.221		
	$I_{\theta-D_{\eta}I_{\theta}}$	0.309	0.221		

2017), where small values indicate better fit. The LOO information criterion has been shown to be superior to other commonly used methods of IRT model comparisons such as the AIC or DIC (Fujimoto & Falk, 2023; Y. Luo & Al-Harbi, 2017).<sup>5</sup>

## Results

The comparison of the co-occurring model ( $I_{\theta-D_{\eta}I_{\theta}}$ ) with the trait-ERS ( $I_{\theta-D_{\eta}}$ ) and trait-trait ( $I_{\theta-I_{\theta}}$ ) models of sequential processes in terms of recovering person and item parameters is summarized in Table 1 and Table 2, respectively. In general, if the co-occurring model was used to generate the data, the model itself provided considerably lower MABs of estimated parameters than both unidimensional models of sequential processes. In contrast, if one of the models of sequential processes was used to generate the data, the co-occurring model yielded MABs of almost equal size. Thus, the co-occurring model successfully adapted to data sets generated by models of sequential processes nested within it, whereas applying such to co-occurring data led to poor parameter recovery.

The evaluation of the out-of-sample model fit (see Table 3) supports these findings: If the

<sup>5</sup>Additional analyses are provided in the OSF project, including recovery by root mean square error and correlation of generated and estimated parameters as well as model fit by the widely applicable information criterion (WAIC; Watanabe, 2010).

data were generated under the co-occurring model, this was superior to the models of sequential processes and was selected as the best-fitting model in all replications. The average LOO values of the models of sequential processes were considerably larger (i.e., indicating worse fit), also when the uncertainty of the LOO estimates is taken into account. In contrast, the differences in the model fit for sequential data were rather small: The respective true model was still selected as the best-fitting model in a large proportion of replications, though in some replications, the co-occurring model provided a better fit. Further, the average LOO values of the co-occurring model were only slightly larger than the values of the respective true model, and such differences were small compared to the standard errors of the LOO estimates. This suggests that the co-occurring model adapted comparably well to the data of sequential processes and successfully captured the restrictions of models nested within it.

Altogether, the simulation study showed that the new DI-MIRT parameterization of IRTree pseudo-items is beneficial for the analysis of item response data and should be preferred over traditional IRTree models of sequential processes. Analyzing data with co-occurring processes under the assumption of sequential processing, that is, ignoring one of two processes, led to poorer model fit and larger errors of estimated parameters. In contrast, there were hardly any negative effects of applying the co-occurring model to data generated by more parsimonious sequential ones. The higher-parameterized co-occurring model entailed greater flexibility and

**Table 2***Recovery of Item Parameters by MAB*

Gen.	Analysis	$\tau$		$\delta$		$\lambda_1$ (Agree. $I_\theta$ )		$\alpha_1$ (Int. $D_\eta$ )		$\lambda_2$ (Int. $I_\theta$ )	
		<i>N</i> 500	<i>N</i> 1000	<i>N</i> 500	<i>N</i> 1000	<i>N</i> 500	<i>N</i> 1000	<i>N</i> 500	<i>N</i> 1000	<i>N</i> 500	<i>N</i> 1000
$I_\theta$ - $D_\eta$ $I_\theta$	$I_\theta$ - $D_\eta$ $I_\theta$	0.320	0.257	0.260	0.209	0.133	0.094	0.108	0.081	0.126	0.095
	$I_\theta$ - $D_\eta$	1.260	1.230	0.738	0.568	0.188	0.132	0.166	0.163		
	$I_\theta$ - $I_\theta$	0.505	0.470	0.426	0.353	0.149	0.109			0.252	0.241
$I_\theta$ - $D_\eta$	$I_\theta$ - $D_\eta$	0.275	0.206	0.719	0.584	0.192	0.135	0.101	0.070		
	$I_\theta$ - $D_\eta$ $I_\theta$	0.286	0.200	0.490	0.412	0.191	0.134	0.104	0.070		
$I_\theta$ - $I_\theta$	$I_\theta$ - $I_\theta$	0.290	0.227	0.233	0.184	0.126	0.089			0.113	0.082
	$I_\theta$ - $D_\eta$ $I_\theta$	0.300	0.232	0.225	0.180	0.126	0.089			0.119	0.085

**Table 3***Model Comparison by LOO*

Gen.	Analysis	$N500, I10$			$N500, I20$			$N1000, I10$			$N1000, I20$		
		LOO	<i>SE</i>	Prop.	LOO	<i>SE</i>	Prop.	LOO	<i>SE</i>	Prop.	LOO	<i>SE</i>	Prop.
$I_{\theta}-D_{\eta}I_{\theta}$	$I_{\theta}-D_{\eta}I_{\theta}$	14321	99	1.00	27743	144	1.00	28517	139	1.00	54889	205	1.00
	$I_{\theta}-D_{\eta}$	15093	94	0.00	29503	140	0.00	30131	132	0.00	58416	199	0.00
	$I_{\theta}-I_{\theta}$	15304	90	0.00	30134	131	0.00	30502	127	0.00	59561	187	0.00
$I_{\theta}-D_{\eta}$	$I_{\theta}-D_{\eta}$	15146	93	0.95	29597	137	1.00	30158	131	0.98	58812	194	0.99
	$I_{\theta}-D_{\eta}I_{\theta}$	15164	93	0.05	29623	138	0.00	30174	132	0.02	58837	194	0.01
$I_{\theta}-I_{\theta}$	$I_{\theta}-I_{\theta}$	14924	91	0.86	29127	134	0.99	29690	128	0.90	58141	188	1.00
	$I_{\theta}-D_{\eta}I_{\theta}$	14934	92	0.14	29156	135	0.01	29702	129	0.10	58170	189	0.00

*Note.* LOO = Average LOO value across replications. *SE* = Average standard error of LOO estimate across replications. Prop. = Proportion of replications with smallest LOO.

was better able to compensate for possible misspecification.

### Empirical Applications

To illustrate the benefits of the general IRTree framework with DI-MIRT parameterization under real-world conditions, two empirical applications of co-occurring dominance and ideal point processes are presented. They relate to the two models described in section IRTree Models of Co-Occurring Processes.<sup>6</sup>

#### Response Style Analysis in Ideal Point Items

In the first application example, the co-occurring IRTree model of RS analysis in ideal point items was applied to a data set consisting of item responses of  $N = 1505$  participants to  $I = 15$  items measuring attitudes toward sexual practices by a subscale of the National Health and Social Life Survey (Laumann et al., 1992)<sup>7</sup>. The items were rated on a four-point scale, with categories "not at all appealing", "not appealing", "somewhat appealing", and "very appealing".

<sup>6</sup>The Stan code of such models can be found in the OSF project.

<sup>7</sup>The data are provided here: <https://www.icpsr.umich.edu/web/HMCA/studies/6647>

The ordinal categories  $Y_{vi} \in \{0, \dots, 3\}$  were decomposed into two sub-decisions of agreement and intensity as defined in Table 4. With the exception that the intensity pseudo-items had two instead of three options, the co-occurring model as described in the section Response Style Analysis in Ideal Point Items was applied.

The data set was previously used as an application example by Jin et al. (2022; using a subset of the data with 1498 respondents), and the authors showed that a trait-ERS model of sequential processes ( $I_{\theta-D_{\eta}}$ ) fitted the data better than an ordinal ideal point model ignoring RS, and than IRTree models under the dominance assumption. Here we analyze the data further and examine whether the co-occurring model ( $I_{\theta-D_{\eta}I_{\theta}}$ ) fits the data even better, which would indicate that the decisions about the intensity of responses were additionally influenced by the ideal point trait.

We used the same analysis scheme as described in the simulation study. The initial values for identifying the orientation of the latent continuum were set on the basis of the estimates reported by Jin et al. (2022). To this end, constrained models without setting initial values were fitted each. If the order of estimated item locations was inverted compared to the results of the previous study, the initial values of item locations and substantive trait levels were set as minus one times the estimates of the constrained model (this was the case for the co-occurring model). Otherwise, the estimates of the constrained model were directly used as the initial values (this was the case for the model of sequential processes). Both models converged with  $\widehat{R} < 1.05$ .

Model comparisons clearly showed that indeed, the new co-occurring model fitted the data well, since the LOO information criterion of this model ( $LOO = 33726$ ) was substantially smaller than the one of the model of sequential processes ( $LOO = 36537$ ). Thus, both a trait-based ideal point process and an ERS-based dominance process were involved in the respondents' decisions regarding the intensity of their responses. This result provides empirical support for multidimensional sub-decisions and underlines the importance of modeling co-occurring processes in IRTree pseudo-items in addition to sequential ones.

In light of this finding, we further analyzed the discrimination parameters of the co-occurring model (see Table 5), as these provide information about the relative importance of each of the processes for the two sub-decisions. Overall, the estimates of the co-occurring model were

**Table 4***Definition of Pseudo-Items for Responses to Four-Point**Rating Items*

Pseudo-item	Ordinal category			
	0	1	2	3
$X_{1vi}$ (Agreement)	0	0	1	1
$X_{2vi}$ (Intensity   agree)	–	–	0	1
$X_{3vi}$ (Intensity   disagree)	1	0	–	–

consistent with previous studies on co-occurring dominance processes (e.g., Meiser et al., 2019; Merhof & Meiser, 2023): Firstly, the discriminating power of trait-based agreement was larger than that of trait-based intensity judgments, suggesting higher importance of the trait for global agreement compared to fine-grained decisions among agreement or disagreement categories. In addition, trait-based and ERS-based processes appear to have similar impacts on intensity decisions, as indicated by discrimination parameters of comparable size. Moreover, the item-specific discrimination parameters of trait-based agreement correlated positively with those of trait-based intensity, but not with those of ERS-based intensity judgments. This correlation pattern also seems reasonable since decisions made on the basis of one and the same personal characteristic, in this case the substantive trait, can be assumed to be interrelated within a single item, whereas trait-based and ERS-based decisions are considered independent processes. This interpretation is supported by the fact that the person variables (i.e., the trait and ERS factors) were only weakly correlated ( $\hat{r}(\theta, \eta) = -.16$ ).

### Middle Categories in Dominance Items

The second empirical example of the DI-MIRT parameterization for co-occurring processes relates to modeling middle categories in dominance items. Response time (RT) data were included in the analysis of item responses in order to put the construct validity of estimated IRTree model parameters to the test. To this end, we examined whether RTs were sensitive to the psychological processes reflected by specific parameters of the IRTree model and changed

**Table 5**

*Estimated Discrimination Parameters of the Co-Occurring  $I_\theta$ - $D_\eta I_\theta$  Model Fitted to Empirical Data*

Parameter	Mean	Min	Max	Correlation	
				$\alpha_i$	$\lambda_{2i}$
$\lambda_{1i}$ (Agreement $I_\theta$ )	1.553	0.913	2.756	-0.015	0.548
$\alpha_i$ (Intensity $D_\eta$ )	1.338	0.480	2.909		0.383
$\lambda_{2i}$ (Intensity $I_\theta$ )	1.119	0.764	1.517		

as hypothesized, which would corroborate reasonable substantive interpretations of the IRT estimates and a meaningful model. We used an empirical data set collected by Fladerer et al. (2021) and Henninger and Plieninger (2020), consisting of item responses and corresponding RTs of  $N = 786$  participants to two questionnaires, the Identity Leadership Inventory ( $I = 14$ ) and a scale of Social Identification ( $I = 6$ )<sup>8</sup>. The items were rated on a five-point rating scale, and the ordinal categories were decomposed into three sub-decisions of midscale responding, agreement, and extreme responding as defined in Figure 6.

In an initial analysis, for which only the item response data were used, the LOO model fit of a co-occurring model described in the section Middle Categories in Dominance Items ( $D_{\eta_1} I_\theta$ - $D_\theta$ - $D_{\eta_2} D_\theta$ ) was assessed. The model assumes that the sub-decisions of midscale and extreme responding depend on the substantive trait plus the MRS or ERS, respectively. Note that although the items are considered dominance items, the substantive trait is modeled as an ideal point process in the midscale sub-decision, as non-midscale categories are expected to be more likely for respondents whose trait levels are more strongly deviating from the item in either an upward or downward direction. Thus, a dominance and an ideal point process co-occur in the midscale pseudo-item, whereas two dominance processes co-occur in the pseudo-items of extreme responding. In order to test our assumption of trait-based responding being an ideal point process in the midscale pseudo-item, we compared this model with an alternative model in which all processes were considered dominance processes ( $D_{\eta_1} D_\theta$ - $D_\theta$ - $D_{\eta_2} D_\theta$ ). In a second

<sup>8</sup>The data are made available by the original authors here: <https://osf.io/gqb4y/>



alternative model of sequential processes ( $D_{\eta_1}-D_{\theta}-D_{\eta_2}$ ), only the agreement sub-decision was defined as dependent on the trait, so midscale and extreme responding were parameterized by unidimensional models of the respective RS.

All three models converged with  $\widehat{R} < 1.05$ . Note that even though an ideal point process was modeled in the co-occurring model, it was not necessary to set initial values for identifying the orientation of the latent continuum. This is because the item locations were set equal to the item-specific thresholds of agreement (see Equation 15), which in turn are inherently identified by the dominance modeling.

The model comparisons revealed that the proposed model of co-occurring processes yielded a considerably better fit ( $LOO = 30656$ ) than the alternative model with dominance processes ( $LOO = 31923$ ), demonstrating that trait-based midscale responding was indeed better described by the ideal point rationale. Further, the model also provided a better fit than the model of sequential processes ( $LOO = 32333$ ), indicating that respondents used both the trait and a RS for the decisions of midscale and extreme responding. The estimated discrimination parameters of the co-occurring model supported this assumption, as they were of substantial size for all processes in all sub-decisions (see Table 6).

A subsequent analysis targeted at the construct validation of the co-occurring model addressed not only the item response data, but additionally the item-level RTs, and both kinds of data were included in a joint model. The item responses were modeled by the co-occurring  $D_{\eta_1}I_{\theta}-D_{\theta}-D_{\eta_2}D_{\theta}$  IRTree model. The RTs were log-transformed and analyzed by linear mixed modeling, whereby the predictor variables included IRTree model parameters. Such a joint model allowed to test whether the RTs were sensitive to specific IRTree parameters and changed according to theory-driven hypotheses, which in turn would suggest that the model produced reasonable estimates.

Our hypotheses on how the parameters of the co-occurring IRTree model should affect the RTs were twofold: Firstly, we assumed that the more the item responses were based on the respondents' individual RS levels, the faster they should be given. The literature suggests that fast responses are associated with low motivation, low data quality, and insufficient effort responding (Bowling et al., 2021; Callegaro et al., 2009; Zhang & Conrad, 2014). Since RS-

**Table 6**

*Estimated Discrimination Parameters of the Co-Occurring  $D_{\eta_1} I_{\theta} - D_{\theta} - D_{\eta_2} D_{\theta}$  Model Fitted to Empirical Data*

Parameter	Identity leadership			Social identification		
	Mean	Min	Max	Mean	Min	Max
$\alpha_{1i}$ (Midscale $D_{\eta_1}$ )	0.936	0.646	1.210	0.694	0.527	0.921
$\lambda_i$ (Midscale $I_{\theta}$ )	1.819	0.691	2.527	1.373	1.178	1.747
$\alpha_{2i}$ (Agreement $D_{\theta}$ )	3.076	1.739	4.898	1.982	1.074	2.830
$\alpha_{3i}$ (Extreme $D_{\eta_2}$ )	1.642	1.076	2.086	0.879	0.642	1.097
$\alpha_{4i}$ (Extreme $D_{\theta}$ )	1.351	0.661	2.205	1.398	0.663	2.233

based responding is a heuristic process requiring less cognitive effort than accurate trait-based responding (Krosnick, 1991; Podsakoff et al., 2012), selecting response categories that match the individual RS should correspond to short RTs. This hypothesis was also investigated by Henninger and Plieninger (2020) in their original work using the data we reanalyzed, and indeed, they found that responses which matched the person-specific RS were given faster. However, they used a two-step approach and obtained estimates of RS levels by an aggregation procedure of dichotomous responses style indicators (i.e., the respondents' ERS and MRS levels were computed based on the information on whether the given responses were extreme versus non-extreme and midscale versus non-midscale, respectively). Here in contrast, we analyzed the data by the joint model, in which the RS levels were estimated by the co-occurring IRTree model in a one-step approach. Nonetheless, we expected to find similar effects, namely shorter RTs for responses that matched the preferred categories.

Most importantly, our second assumption concerned the ideal point modeling of trait-based midscale responding, and we expected that large distances between the respondents' trait levels and the items' locations would result in fast responses. This reasoning relates to a hypothesis that has been frequently described in the literature under terms such as speed-distance or distance-difficulty hypothesis (e.g., Ferrando & Lorenzo-Seva, 2007; McIntyre, 2011; Ulitzsch et al., 2022). It states that a large person-item distance on the latent trait continuum evokes high certainty, which in turn, should be reflected in clear-cut (compared to moderate) responses

and shorter RTs. Part of this hypothesis was also already tested and supported by Henninger and Plieninger (2020), as they found that selecting the middle category was associated with longer RTs, indicating that such responses were related to uncertainty. However, we further analyzed whether RTs were not only dependent on the selected rating category per se (e.g., whether an extreme or midscale category was selected), but additionally affected by the distance of latent person and item locations. We defined the respondents' locations as the estimated substantive trait levels obtained by the IRTree model and the items' locations as estimated difficulty parameters of the agreement sub-decision. The agreement difficulty parameter was used, as it determines for which trait levels the general attitude toward the item statement is rather positive or negative, and thus marks the point of maximal uncertainty. Note that this person-item distance is also part of the IRTree pseudo-item of midscale responding: In this pseudo-item, the substantive trait levels represent the ideal points of the respondents with respect to the midscale sub-decision, and the item locations are set equal to the agreement difficulty parameters (see  $X_{1vi}$  in Equation 15). Therefore, the person-item distance is assumed to affect both the RTs (a higher distance should result in shorter RTs) and the probability of a trait-based selection of middle categories (a higher distance should be associated with a lower probability).

The linear mixed model for predicting the log-transformed RTs of a response  $r$  given by person  $v$  to item  $i$  is defined by

$$\begin{aligned} \log(RT_{rvi}) = & \gamma_{000} + \gamma_{1v0} \times X_{1vi} + \gamma_{2v0} \times X_{3vi} + \gamma_{2v0} \times X_{4vi} + \\ & \gamma_{011} \times |\theta_v - \beta_{2i}| + u_{0v0} + u_{00i} + \epsilon_{rvi} \end{aligned} \quad (16)$$

with

$$\begin{aligned} \gamma_{1v0} &= \gamma_{100} + \gamma_{110} \times \eta_{1v}, \\ \gamma_{2v0} &= \gamma_{200} + \gamma_{220} \times \eta_{2v}. \end{aligned} \quad (17)$$

The predictors  $X_{hvi}$  refer to the IRTree pseudo-items as defined in Figure 6 and indicate whether a given response was the middle category ( $X_{1vi}$ ) or one of the extreme categories ( $X_{3vi}$  or  $X_{4vi}$ ). As those predictors are manifest observations, they do not relate to the IRTree model and were merely included as control variables. In addition, random person and item effects ( $u_{0v0}$  and  $u_{00i}$ , respectively) were included to account for the fact that some respondents are generally faster than others and that some items are faster to respond to than others. Predictors resulting from the IRTree model and referring to the substantial hypotheses were the MRS levels  $\eta_{1v}$ , the ERS levels  $\eta_{2v}$ , and the person-item distances  $|\theta_v - \beta_{2i}|$ . The effect of RS levels matching a given response (hypothesis 1) was captured by  $\gamma_{110}$  and  $\gamma_{220}$ . The effect of the person-item distance (hypothesis 2) was captured by  $\gamma_{011}$ .

The results of our analysis using the joint model corroborated both hypotheses (see Table 7): Firstly, we found that heuristic, RS-based responding was related to short RTs. Both for midscale and extreme responding, a match of individual preferences with the selected category reduced the predicted RT in the mixed model (Person x Response level). Further, selecting the middle category was associated with on average longer RTs and extreme responses with shorter RTs (Response level). Thus, the closer the selected category was to the middle of the scale, the more time respondents needed, which indicates that such decisions were related to higher uncertainty. Importantly, a larger person-item distance was associated with shorter RTs in addition to this effect of the selected category (Person x Item level). Therefore, the speed-distance hypothesis was supported not only at the level of manifest response categories, but also at the level of latent locations estimated by the IRTree model. This result is original evidence that the absolute person-item distance, regardless of the direction, influenced both the time it took respondents to choose a category and the probability of selecting the middle category (see estimates of  $\lambda_i$  in Table 6). The direction of this distance, however, affected the probability to agree or disagree with the item (see estimates of  $\alpha_{2i}$  in Table 6).

Altogether, the estimates produced by the co-occurring IRTree model affected the RTs in accordance with our theory-driven hypotheses. This suggests that the new DI-MIRT model appropriately captured the co-occurring response processes, and corroborates the construct validity of the applied IRTree model.

**Table 7***Estimated Coefficients of the Linear Mixed Model Predicting Log-Transformed RTs*

Level	Predictor	Coefficient	Estimate	95 %-credible interval
Response	Middle category	$\gamma_{100}$	0.041	[0.016; 0.067]
	Extreme category	$\gamma_{200}$	-0.068	[-0.093; -0.042]
Person x Response	MRS level x middle cat.	$\gamma_{110}$	-0.116	[-0.152; -0.079]
	ERS level x extreme cat.	$\gamma_{220}$	-0.127	[-0.155; -0.098]
Person x Item	Person-item distance	$\gamma_{011}$	-0.086	[-0.105; -0.067]

### Conclusion

The present article introduced a general IRTree framework for modeling multidimensional response processes with dominance and ideal point item response functions (IRFs). Such response processes (e.g., responding based on the substantive trait or based on response styles; RS) can be defined to be involved in item responding both sequentially across sub-decisions and as co-occurring processes within sub-decisions. Unlike sequential multidimensionality, which can be implemented using existing IRT modeling (see Jin et al., 2022), co-occurring response processes have previously been limited exclusively to dominance models (e.g., Alagöz & Meiser, 2023; Jeon & De Boeck, 2016; Meiser et al., 2019; Merhof & Meiser, 2023; von Davier & Khorramdel, 2013). Therefore, we developed a new multidimensional IRT model of co-occurring dominance and ideal point processes (DI-MIRT model), with which multiple dominance processes, multiple ideal point processes, as well as a combination of both can be included in IRTree pseudo-items in a consistent way. The proposed DI-MIRT parameterization expands the toolbox of IRTree models and thereby opens up new application areas for this model class. A wide range of theoretical assumptions about the cognitive processing during item responding can be specified within the new general IRTree framework, in which different components can be flexibly combined in the sense of a modular system. Independent choices can be made regarding the decomposition of ordinal responses into sub-decisions, the assignment of response processes to the sub-decisions, and the selection of IRFs for the individual processes. Such components can be freely defined and adapted to the research question and the data at hand.

A simulation study demonstrated that the proposed IRTree framework with DI-MIRT parameterization of pseudo-items accurately captured co-occurring processes and recovered the person and item parameters well. Furthermore, it also showed good parameter recovery in the case of over-parameterization, that is, when applied to data generated under IRTree models in which multiple response processes were only involved sequentially across pseudo-items. In contrast, if one of the co-occurring processes was ignored and a parsimonious IRTree model of sequential processes was falsely applied, larger errors of estimated parameters and poorer model fit resulted. These findings indicate that multidimensional pseudo-items should be preferred over unidimensional ones, wherever this seems reasonable from a theoretical point of view. The DI-MIRT model facilitates this for both dominance and ideal point processes and goes beyond previous MIRT models, which were limited to specific kinds of processes (Bolt & Johnson, 2009; Bolt & Newton, 2011; Falk & Cai, 2016; Henninger & Meiser, 2020; Javaras & Ripley, 2007; Jin & Wang, 2014; Liu & Wang, 2019).

Two empirical examples further demonstrated the advantage of the new IRTree parameterization under realistic conditions. In the first example, a co-occurring model was used to analyze the influence of ERS on responding to ideal point items, in which trait-based responding was modeled under the ideal point assumption and ERS-based responding under the dominance assumption. Both kinds of response processes were considered for modeling the sub-decisions of extreme versus non-extreme responding by using the DI-MIRT parameterization, and the results showed that indeed the trait and the ERS co-occurred in such decisions, which is why ignoring one of the two processes led to a substantially worse model fit. The exemplary IRTree model used for analyzing this data set can be easily adapted to other applications of co-occurring trait and RS effects with differently structured trees, different sub-decisions, or other RS. Further extensions are also conceivable with respect to additional influences apart from RS, such as socially desirable responding, which likewise follows the dominance rationale. Thereby, the default assumption in the literature that traits are dominance processes can be challenged and compared to the alternative ideal point assumption, while taking further response processes into account. Such investigations seem promising, as the previous research has shown that even if items were constructed as dominance items, ideal point models may better describe the response

behavior of respondents (Dragow et al., 2010).

The second empirical example of this article made use of the DI-MIRT parameterization for examining the respondents' use of middle categories. It was shown that the substantive trait as well as the MRS influenced such decisions, and that multidimensional pseudo-items fitted the data better than the unidimensional ones. The additional analysis of response time data further supported the construct validity of the estimated DI-MIRT parameters, as the relation of parameters and response times was in line with the theory-driven hypotheses. Moreover, the model used in this application demonstrated that response processes do not necessarily adhere to fixed IRFs (i.e., are inherently dominance or ideal point processes), but that it may be beneficial to assign different IRFs to one and the same process across IRTree pseudo-items: Although the items were considered dominance items, meaning that trait-based agreement was modeled as dominance process, trait-based midscale responding was defined as an ideal point process. This choice of IRFs reflected our hypothesis that midscale responding was unlikely for both very high and very low trait levels in relation to the item location, which was indeed supported by the data. Such a varying assignment of IRFs could also be useful for other research questions. For instance, one could assume that the respondents first decide on whether the item generally fits their own attitude (i.e., trait-based agreement follows the ideal point rationale), but subsequently respond according to a more-is-better principle (i.e., fine-grained sub-decisions are reflected by the dominance rationale).

Furthermore, co-occurring dominance and ideal point response processes may exist outside self-reported rating data, such as in the field of educational research and ability measurement: For example, the performance in low-stakes assessments might not exclusively be the result of a dominance process with the probability of correct responding being monotonously increasing with higher ability levels. Instead, respondents with very high ability levels may not feel sufficiently challenged and respond with a somewhat lower effort than others, which may result in a lower-than-expected performance. In such cases, a combination of ideal point and dominance IRFs might be appropriate, resulting a steep increase in expected performance from low to high trait levels and a slight decrease for even higher levels. As responding to performance items can usually not be decomposed into different sub-decisions, and as the responses

to such items are typically coded as correct or incorrect, the DI-MIRT model could be used as an ordinal or dichotomous model without implementing it within the IRTree framework. A similar dichotomous DI-MIRT model might be suitable for investigating missing responses in performance tests, where a higher-than-expected number of item omissions could likewise occur for respondents who are not sufficiently challenged.

In addition, a DI-MIRT parameterization of IRTree models could be used for modeling missing responses in Likert-type rating data, for instance, as an extension of the missing model introduced by Jeon and De Boeck (2016). The authors proposed an IRTree model in which respondents first decide on whether they wanted to omit the item based on their omission propensity, and then optionally answer the item and chose one of the available categories based on the substantive trait. As a further development to the original model, the omission sub-decision could be given a two-dimensional parameterization of both the omission propensity and the trait. While the omission propensity can be considered a dominance process, the ideal point assumption seems reasonable for the trait-based response process: Mainly respondents who have moderately high trait levels in relation to the item are expected to omit the response, whereas respondents with very high or low trait levels are unlikely to skip the item, as they should have clear-cut opinions. Thus, the ideal point trait could be combined with the dominance omission propensity in the omission sub-decision using the DI-MIRT model, similar to the modeling approach of middle categories in the present article.

Another possible application of the proposed DI-MIRT model (within or outside the IRTree framework) is the co-occurrence of two ideal point response processes. For instance, a bifactor model may reflect the factor structure of a questionnaire with several interrelated sub-scales. If both the general factor and the specific factors are assumed to follow the ideal point rationale, multidimensional modeling of several ideal point processes would be required, which can be achieved by the DI-MIRT model.

A limitation of the proposed DI-MIRT approach is the assumption that the same composition of response processes holds for all respondents.<sup>9</sup> It seems likely that individuals differ

---

<sup>9</sup>We thank an anonymous reviewer for pointing this out.



in what response processes they use to what extent in empirical data, especially if the circumstances of the data collection vary (e.g., because the respondents' motivation or perceived time pressure differ). As a result, some respondents may derive their answers solely based on their substantive trait, while others may additionally use RS. Further, respondents may also differ in how they perceive the item statements and the rating scale, which could lead to some respondents applying trait-based responding in a dominance way, while others may rather respond in an ideal point fashion. The model proposed here cannot account for such heterogeneity between respondents and instead will reveal the average group-level response behavior. More detailed insights about the item response process would be obtained if interindividual differences were considered, for example, by extending the DI-MIRT model by a person mixture. Though such an approach appears very promising from the theoretical perspective, future studies would be needed to evaluate the practical feasibility of estimating group-specific parameters in the DI-MIRT framework.

A further limitation of the present work is that we investigated the co-occurrence of only two response processes at most. We considered this as a plausible assumption within the IRTree framework since more than two processes nevertheless can contribute to item responding across sub-decisions. Moreover, this ensured that the complexity of the models was kept at a reasonable level. Although the DI-MIRT approach comprises a wide range of potentially very complex models, heavily parameterized models including many response processes should be used with caution, as it may become impossible to disentangle and interpret the defined processes. Instead, researchers should specify their models based on theoretical considerations and compare models with increasing complexity against each other in order to select a well-fitting but interpretable model. With this in mind, the DI-MIRT model introduced here offers a versatile approach for psychometricians in various fields of research and practice.

## References

- Alagöz, E., & Meiser, T. (2023). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, Advance online publication. <https://doi.org/10.1177/00131644231206765>
- Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. John Wiley & Sons.
- Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, *19*(3), 269–290. <https://doi.org/10.1177/014662169501900306>
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, *17*(3), 253–276. <https://doi.org/10.1177/014662169301700307>
- Baumgartner, H., & Steenkamp, J.-B. E. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. <https://doi.org/10.1509/jmkr.38.2.143.18840>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. <https://doi.org/10.1007/BF02291411>
- Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. <https://doi.org/10.1037/a0028111>
- Böckenholt, U. (2017). Measuring response styles in Likert items. *Psychological Methods*, *22*(1), 69–83. <https://doi.org/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 and Statistical Psychology*, *70*(1), 159–181. <https://doi.org/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*(5), 335–352. <https://doi.org/10.1177/0146621608329891>
- Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814–833. <https://doi.org/10.1177/0013164410388411>
- Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*. <https://doi.org/10.1177/10944281211056520>
- 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 of Sociological Methodology/Bulletin de Méthodologie Sociologique, 103*(1), 5–25. <https://doi.org/10.1177/075910630910300103>
- Coombs, H. C. (1964). *A theory of data*. John Wiley.
- Cui, W. (2008). *The multidimensional generalized graded unfolding model for assessment of change across repeated measures*. [Doctoral dissertation, University of Maryland]. College Park ProQuest Dissertations Publishing.
- De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software, 48*(1), 1–28. <https://doi.org/10.18637/jss.v048.c01>
- Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3*(4), 465–476. <https://doi.org/10.1111/j.1754-9434.2010.01273.x>
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods, 21*(3), 328–347. <https://doi.org/10.1037/met0000059>
- 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*(6), 525–543. <https://doi.org/10.1177/0146621606295197>

- Fladerer, M. P., Kugler, S., & Kunze, L. G. (2021). An exploration of co-workers' group identification as moderator of the leadership-health link. *Small Group Research, 52*(6), 708–737. <https://doi.org/10.1177/10464964211007562>
- Fujimoto, K. A., & Falk, C. F. (2023). The accuracy of Bayesian model fit indices in selecting among multidimensional item response theory models. *Educational and Psychological Measurement. https://doi.org/10.1177/00131644231165520*
- Gabry, J., Češnovar, R., & Johnson, A. (2023). Cmdstanr: R interface to CmdStan.
- Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods, 25*(5), 560–576. <https://doi.org/10.1037/met0000249>
- Henninger, M., & Plieninger, H. (2020). Different styles, different times: How response times can inform our knowledge about the response process in rating scale measurement. *Assessment, 28*(5), 1301–1319. <https://doi.org/10.1177/1073191119900003>
- Javaras, K. N., & Ripley, B. D. (2007). An “unfolding” latent variable model for Likert attitude data. *Journal of the American Statistical Association, 102*(478), 454–463. <https://doi.org/10.1198/016214506000000960>
- Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods, 48*(3), 1070–1085. <https://doi.org/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*(1), 116–138. <https://doi.org/10.1177/0013164413498876>
- Jin, K.-Y., Wu, Y.-J., & Chen, H.-F. (2022). A new multiprocess IRT model with ideal points for Likert-type items. *Journal of Educational and Behavioral Statistics, 47*(3), 297–321. <https://doi.org/10.3102/10769986211057160>
- Kalton, G., Roberts, J., & Holt, D. (1980). The effects of offering a middle response option with opinion questions. *The Statistician, 29*(1), 65. <https://doi.org/10.2307/2987495>

- 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*(2), 161–177. <https://doi.org/10.1080/00273171.2013.866536>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. <https://doi.org/10.1002/acp.2350050305>
- Laumann, E. O., Gagnon, J. H., Michael, R. T., & Michaels, S. (1992). *National health and social life survey, 1992*. Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR06647.v2>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5–53.
- Liu, C.-W., & Wang, W.-C. (2016). Unfolding IRT models for Likert-type items with a don't know option. *Applied Psychological Measurement, 40*(7), 517–533. <https://doi.org/10.1177/0146621616664047>
- Liu, C.-W., & Wang, W.-C. (2019). A general unfolding IRT model for multiple response styles. *Applied Psychological Measurement, 43*(3), 195–210. <https://doi.org/10.1177/0146621618762743>
- Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. *Journal of Mathematical Psychology, 42*(4), 400–417. <https://doi.org/10.1006/jmps.1998.1206>
- Luo, Y., & Al-Harbi, K. (2017). Performances of LOO and WAIC as IRT model selection methods. *Psychological Test and Assessment Modeling, 59*(2), 183–205.
- 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*(5), 597–602. <https://doi.org/10.1016/j.paid.2010.12.001>
- 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 and Statistical Psychology, 72*(3), 501–516. <https://doi.org/10.1111/bmsp.12158>

- Merhof, V., & Meiser, T. (2023). Dynamic response strategies: Accounting for response process heterogeneity in IRTree decision nodes. *Psychometrika*, *88*(4), 1354–1380. <https://doi.org/10.1007/s11336-023-09901-0>
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. <https://doi.org/10.1177/014662169201600206>
- Nowlis, S. M., Kahn, B. E., & Dhar, R. (2002). Coping with ambivalence: The effect of removing a neutral option on consumer attitude and preference judgments. *Journal of Consumer Research*, *29*(3), 319–334. <https://doi.org/10.1086/344431>
- Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. <https://doi.org/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*(1), 539–569. <https://doi.org/10.1146/annurev-psych-120710-100452>
- R Core Team. (2023). R: A language and environment for statistical computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3–32. <https://doi.org/10.1177/01466216000241001>
- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement*, *20*(3), 231–255. <https://doi.org/10.1177/014662169602000305>
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(Suppl 1), 1–97. <https://doi.org/10.1007/BF03372160>
- Stan Development Team. (2023). Stan modeling language users guide and reference manual, version 2.33. <https://mc-stan.org>

- Sturgis, P., Roberts, C., & Smith, P. (2014). Middle alternatives revisited. *Sociological Methods and Research*, *43*(1), 15–38. <https://doi.org/10.1177/0049124112452527>
- Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering (Ed.), *Handbook of polytomous item response theory models*. Routledge.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. <https://doi.org/10.1007/BF02295596>
- Thissen-Roe, A., & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547. <https://doi.org/10.3102/1076998613481500>
- Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554. <https://doi.org/10.1086/214483>
- Tijmstra, J., & Bolsinova, M. (in press). Modeling within- and between-person differences in the use of the middle category in Likert scales. *Applied Psychological Measurement*.
- Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. <https://doi.org/10.3758/s13428-017-0997-0>
- Ullrich, E., Pohl, S., Khorrandel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. <https://doi.org/10.1007/s11336-021-09817-7>
- van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement*, *18*(2), 97–110. <https://doi.org/10.1177/014662169401800201>
- 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*(2), 195–217. <https://doi.org/10.1093/ijpor/eds021>
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>

- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved  $\widehat{R}$  for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, *16*(2). <https://doi.org/10.1214/20-BA1221>
- von Davier, M., & Khorramdel, L. (2013). Differentiating response styles and construct-related responses: A new IRT approach using bifactor and second-order models. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology* (pp. 463–487). Springer. [https://doi.org/10.1007/978-1-4614-9348-8\\_30](https://doi.org/10.1007/978-1-4614-9348-8_30)
- Wang, W.-C., Liu, C.-W., & Wu, S.-L. (2013). The random-threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement*, *37*(3), 179–200. <https://doi.org/10.1177/0146621612469720>
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.
- Zhang, C., & Conrad, F. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135. <https://doi.org/10.18148/srm/2014.v8i2.5453>



## C Acknowledgements

The journey of my dissertation was an exciting and rewarding chapter of my life, from which I take with me not only the knowledge and skills I gained but also cherished memories and invaluable relationships. Beginning in the shadow of a pandemic, quite alone at my desk at home, it evolved into an array of enriching encounters and learning opportunities, shaping me both professionally and personally. For all these experiences and for everyone who has been a part of my journey, I am very grateful.

First and foremost, I want to express my deepest appreciation to Thorsten Meiser. Thank you for your unwavering support, your encouragement to broaden my horizon, and for the unrestricted freedom to realize my ideas.

I am grateful to the SMiP research training group and all the people who make up this wonderful group for giving me the opportunity to pursue my PhD in such an inspiring and motivating environment. Thank you for enlightening conversations, helpful feedback, and new perspectives. Many thanks to Anke Söllner and Annette Förster for all the well-organized SMiP events and for making SMiP as successful as it is. Most of all, I would like to thank my fellow SMiPsters for the many hours we spent together during and after workshops, retreats, and informal meetings, and for sharing the ups and downs of our PhDs with each other.

I thank my colleagues and friends at our chair, Nils, Emre, Timo, Julia, Marcel, and Susanne, for their company and support. Our lunch breaks, Blitzlichtrunden, conference trips, and bouldering events not only inspired insightful discussions on research but also infused my scientific pursuits with a dimension of fun and delight. I would like to express my special thanks to Fabiola for waking up the social life at our chair from its Covid-hibernation and for always having an open door and listening ear.

I am deeply grateful to Erhan Genç for sparking the joy of research in me, for entrusting me with responsibility, and for preparing me for a PhD in the best possible way.

I thank all my family and friends for being part of my life and for the memories we created over months, years, and decades. I would like to especially thank my parents, Anja and Mark, for unconditionally supporting me in whatever I dreamed of doing and for their encouragement to find out what this would be.

Finally, I would like to thank Felix for being my partner and best friend. Thank you that we share moments of joy, struggles and challenges, adventures and laughter, and sometimes even thoughts.

This dissertation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) to the Research Training Group “Statistical Modeling in Psychology” (GRK 2277).