



Using Response Times for Joint Modeling of Careless Responding and Attentive Response Styles

Esther Ulitzsch 

*IPN—Leibniz Institute for Science and Mathematics Education
Centre for International Student Assessment (ZIB)*

Steffi Pohl 

Freie Universität Berlin

Lale Khorramdel

Boston College

Ulf Kroehne

DIPF—Leibniz Institute for Research and Information in Education

Matthias von Davier

Boston College

Questionnaires are by far the most common tool for measuring noncognitive constructs in psychology and educational sciences. Response bias may pose an additional source of variation between respondents that threatens validity of conclusions drawn from questionnaire data. We present a mixture modeling approach that leverages response time data from computer-administered questionnaires for the joint identification and modeling of two commonly encountered response bias that, so far, have only been modeled separately—careless and insufficient effort responding and response styles (RS) in attentive answering. Using empirical data from the Programme for International Student Assessment 2015 background questionnaire and the case of extreme RS as an example, we illustrate how the proposed approach supports gaining a more nuanced understanding of response behavior as well as how neglecting either type of response bias may impact conclusions on respondents' content trait levels as well as on their displayed response behavior. We further contrast the proposed approach against a more heuristic two-step procedure that first eliminates presumed careless respondents from the data and subsequently applies model-based approaches accommodating RS. To investigate the trustworthiness of results obtained in the empirical application, we conduct a parameter recovery study.

Keywords: *response bias; careless responding; response styles; response times*

Questionnaires are by far the most common tool for measuring noncognitive constructs in psychology and educational sciences. Above and beyond differences in the traits to be measured, response bias may pose an additional source of variation between respondents that threatens validity of conclusions drawn from questionnaire data. In the present article, we focus on two commonly encountered response bias—careless and insufficient effort responding (C/IER) and response styles (RS) in attentive answering. The former refers to behavior shown by respondents approaching the administered items inattentively and choosing responses that do not reflect the trait to be measured, for example, by random responding or straight lining (Meade & Craig, 2012). With the latter, we refer to response behavior shown by respondents whose responses, although stemming from—at least in parts—attentive response processes are confounded with content-irrelevant variability due to differences in category usage and perception, such as midpoint (MRS) or extreme RS (ERS; Böckenholt & Meiser, 2017).

Note that the key characteristic of the employed distinction between C/IER and RS in attentive answering (or attentive RS) is whether or not observed responses still reflect the trait to be measured; that is, whether the targeted response bias induces content-irrelevant variability on the between-item (C/IER) or within-item (attentive RS) level. As such, C/IER does not merely include seemingly random responses but also subsumes response patterns going back to systematic tendencies of respondents to prefer specific response categories (e.g., outer response options) over others—as long as such systematic tendencies are the only driver of how respondents choose response options and the resulting responses are uninformative of the traits to be measured. Likewise, the employed distinction between C/IER and RS in attentive answering still allows for both types of response bias to potentially stem from noneffortful responding. However, while C/IE responses are assumed to not reflect the traits to be measured whatsoever, we assume that responses that are confounded with attentive RS are still reflective of the traits to be measured to some extent. That is, respondents must have invested at least some effort into reading the item and retrieving relevant information. In the case that RS is a manifestation of lowered effort due to, for example, fatigue, respondents may have read and/or processed the items superficially and employed their category preferences as heuristics in choosing a relevant option (see Lyu & Bolt, 2022, for recent model developments aiming to capture such behavior). Besides fatigue, other sources of RS in attentive answering exist, with cultural differences being the most prominent example (Johnson, 2005). That is, in this study, for the sake of simplicity, we employ the terms “attentive” and “inattentive”/“careless” as complementary antonyms, demarcating zero from non-zero attentiveness, rather than as gradable antonyms.

Vast literature exists that focuses on the identification of either type of response bias (see Böckenholt & Meiser, 2017; Henninger & Meiser, 2020, for overviews on RS; and Meade & Craig, 2012; Niessen et al., 2016, for overviews on C/IER). Although both types of response bias are plausible to be present in questionnaire data, so far, no model exists that considers both bias simultaneously. Separating and jointly considering both types of bias is a challenging endeavor since they may both result in rather similar response vectors. Response vectors consisting of the highest response category on all items, for instance, may either go back to a combination of a high content trait level and attentive ERS or to inattentive straight-lining behavior. Nevertheless, C/IER and attentive RS pose markedly different response processes that can be assumed to differ in other behavioral indicators besides mere responses. Response times (RTs) retrievable from computer-administered questionnaires are a prominent and widely employed example for such behavioral indicators, allowing to both separate and better understand different response processes (e.g., Henninger & Plieninger, 2020; Schnipke & Scrams, 1997; Ulitzsch et al., 2020; Wang & Xu, 2015; Wise, 2017). We aim to develop an approach that draws on the additional information contained in RTs for disentangling as well as simultaneously accounting for and investigating C/IER and attentive RS.

In what follows, we first review current, solely response-pattern-based approaches that support investigating and handling C/IER and attentive RS. We then discuss the potential of RTs for better understanding different aspects of response behavior in questionnaire data. Based on these considerations, we present an RT-based mixture modeling approach that combines and extends approaches for different aspects of response behavior and bias. In doing so, the approach distinguishes attentive respondents from those showing C/IER and separates parts of attentive responses going back to differences in trait levels from those due to RS. In an application of the model to data from the Programme for International Student Assessment 2015 (PISA; Organization for Economic Cooperation and Development, 2017) background questionnaire, we illustrate which insights on response behavior can be gained based on the presented approach and contrast it against analyses leaving either C/IER, attentive RS, or both unconsidered. We further contrast the proposed approach against a more heuristic two-step procedure that first eliminates presumed careless respondents from the data and subsequently applies model-based approaches accommodating RS in attentive responding. To investigate the trustworthiness of results obtained in the empirical application, we assess parameter recovery of the approach in a simulation study.

Response-Pattern-Based Approaches for Identifying Response Biases

Careless and Insufficient Effort Responding

Traditional approaches for C/IER commonly employ response-pattern-based indicators for its identification. Examples for such indicators are the long string

index, being constructed by examining the longest sequence of subsequently occurring identical responses for each respondent (Johnson, 2005), the even-odd index, given by the within-person correlation between the responses to odd-numbered and even-numbered items belonging to the same scale, averaged across scales (Curran, 2016; Huang et al., 2012), or Mahalanobis distance, following the rationale that C/IE responses are outliers that deviate from typical response patterns (Curran, 2016; Huang et al., 2012). Exhaustive overviews and discussions of other response-pattern-based indicators are given in Curran (2016), Meade and Craig (2012), and Niessen et al. (2016).

When respondents exceed a predefined threshold on the employed indicator, they are classified as careless. There is an ongoing discussion on how thresholds should be set, as these can heavily impact conclusions on the occurrence of C/IER (e.g., Curran, 2016; Niessen et al., 2016). A further problematic aspect of these response-pattern-based indicators is that each index is tailored to the detection of a different type of C/IER behavior but insensitive to others. The long string index, for instance, is well suited for detecting straight lining but does not detect diagonal lining or random responding. Conversely, the even-odd index is insensitive to straight lining since this results in consistent response patterns (Curran, 2016). Mahalanobis distance, in contrast, can be influenced by too much normality in C/IE responses (arising when respondents randomly choose categories around the midpoint; Curran, 2016). Thus, Mahalanobis distance performs well for detecting uniformly distributed random responses while failing to detect normally distributed random responses (Meade & Craig, 2012). To accommodate this issue, Curran (2016) suggested to combine multiple indicators that are sensitive to different aspects of C/IER in a multiple-hurdle approach, where respondents with extreme values on any of the considered indicators are filtered out in a stepwise procedure. However, Ulitzsch, Pohl, et al. (2021) illustrated that multiple-hurdle approaches, too, are heavily impacted by threshold choices for the employed indicators.

To avoid making assumptions concerning the specific types of C/IER or attentive response patterns, Schroeders et al. (2020) suggested to employ supervised machine learning techniques, with the algorithm being trained on a data set for which it is known which respondents displayed attentive and C/IER behavior, for example, on data stemming from experiments manipulating instructions on how to approach the questionnaire. This approach, however, requires access to an adequate training data set and is based on the assumptions that (a) respondents in the experimental prestudy complied with instructions, for example, provided attentive responses when instructed to do so, and that (b) both attentive and C/IE responses in the data set of interest follow a structure that is comparable to the respective structures in the training data, that is, that respondents being instructed to show C/IER behavior behave in a comparable manner to those displaying C/IER behavior in out-of-lab conditions.

The abovementioned approaches pose two-step approaches, where, in the first step, careless respondents are filtered out, and in the second step, analysis methods of choice, for example, polytomous item response theory (IRT) models, are applied to the cleaned data set. In step two, researchers could in principle also employ models considering attentive RS, thus jointly accounting for C/IER and attentive RS. To the best of our knowledge, such procedures have not yet been evaluated. A potential limitation may be that misclassifications of the method chosen in step one may impact subsequent RS analyses. It is yet not known how this impacts the overall adjustment procedure. As delineated above, under both indicator-based and machine learning approaches as potential tools for step one, misclassifications are likely to occur.

Attentive RS

Approaches for attentive RS have in common that they aim at disentangling parts of the response process going back to differences in trait levels from those due to differences in category usage. In the last decades, a myriad of model-based approaches for RS has been developed. A dominant stream of research conceptualizes RS as person-specific shifts in item parameters of polytomous IRT models such as the (generalized) partial credit model (PCM; Masters, 1982; Samejima, 2016) or the rating scale model (RSM; Andrich, 1978). Henninger and Meiser (2020) provided an overview and generalized framework subordinating different approaches. The authors delineated that current procedures for modeling RS differ in the restrictions they impose on the structure of RS and/or their relationships to the substantive traits. While models that leave the structure of RS unconstrained (e.g., Bolt & Johnson, 2009; Rost, 1991; Wang & Wu, 2011) allow for an exploratory investigation of RS, they commonly assume RS to be independent of the traits to be measured in the sense that person-specific shifts are uncorrelated with the content traits. Models that impose structures on person-specific shifts constrain these to follow patterns that resemble theory-derived RS, such as MRS or ERS (e.g., Bolt & Johnson, 2009; Falk & Cai, 2016; Tutz et al., 2018; Wetzels & Carstensen, 2015). These models commonly allow for respondents with different trait levels to differ in their stylistic tendencies. In presenting this general framework, Henninger and Meiser (2020) explicated the specific types of RS and research questions that can and cannot be investigated using either of the subsumed models and provided guidelines for their application.

Another dominant stream of research on IRTree models for RS aims at decomposing the response process into subsequent, a priori defined subprocesses (see Böckenholt & Meiser, 2017; Jeon & De Boeck, 2016, for overviews). For instance, a response on an outer agreement category may be decomposed in an agreement process, followed by the decision to opt for an outer response option. IRTree models are based on the construction of binary pseudo items, containing information on the outcomes of each of the assumed subprocesses, and

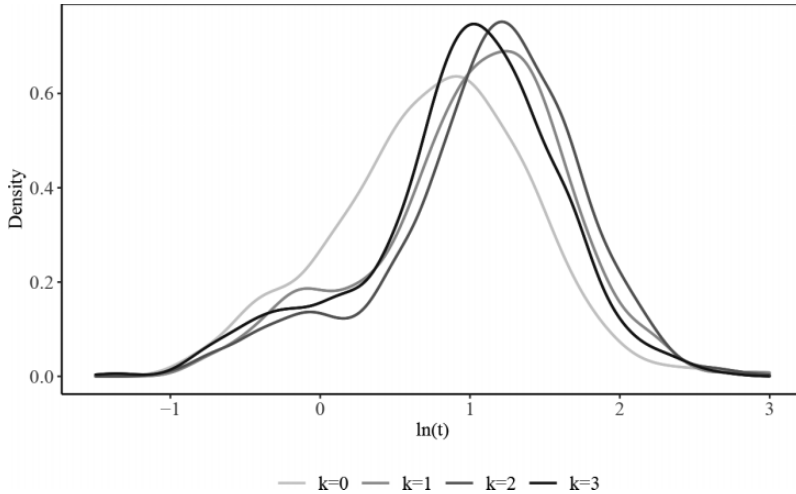


FIGURE 1. Distribution of log response times by response category for Item ST094Q04 of the Programme for International Student Assessment 2015 background questionnaire. Response times were retrieved from raw log events using the finite state machine framework by Kroehne and Goldhammer (2018).

constituting their measurement models. Various extensions of IRTree models exist that allow for a finer-grained investigation of response subprocesses, entailing subprocesses with ordinal and multidimensional decision nodes (Meiser et al., 2019) or mixture extensions that allow for respondents to structurally differ in subprocesses involved for choosing response options (Khorramdel et al., 2019; Kim & Bolt, 2020).

Models for attentive RS pose sophisticated methods for depicting important aspects of response processes. Nevertheless, they commonly assume that all responses reflect the trait to be measured to some degree. This assumption, however, is violated when some respondents display C/IER.

Using Response Times to Investigate Response Biases

Careless and Insufficient Effort Responding

Since inattentive respondents do not invest effort in evaluating the item, retrieving relevant information, and selecting a relevant response, C/IER can be assumed to generally require less time than attentive responding. First traces of multiple processes underlying RTs in computer-administered questionnaires are already revealed in descriptive analyses. Figure 1 displays log RT distributions by response category for a single item from the PISA 2015 background questionnaire completed by students from the German sample. RTs associated

with choosing either of the four response categories show a bimodal shape. We observed similar patterns across various items. In the context of cognitive assessment, such bimodal shapes are commonly assumed to go back to noneffortful rapid guessing behavior and effortful solution behavior, respectively (e.g., Wise, 2017). Comparable conclusions can be drawn in the context of questionnaire data (Kroehne et al., 2019; Ulitzsch et al., 2023).

Various approaches exist that draw on RT information for separating attentive from C/IER behavior. Threshold-based methods aim at identifying thresholds separating RT distributions associated with either behavior, for instance, by making an educated guess on the minimum amount of time required for an attentive response (Huang et al., 2012; Meade & Craig, 2012) or by visual inspection of the RT distribution (Kroehne et al., 2019). Note that RT-based indicators come with the advantage of not entailing presumptions on the specific C/IE response patterns.

The potential of RTs for facilitating the detection of C/IER has also been illustrated by Schroeders et al. (2020) who found that prediction accuracy of their supervised machine learning approach to C/IER could further be improved by jointly considering responses and RTs for classification.

In a similar vein, deliberately incorporating theoretical considerations on response behavior, Ulitzsch, Pohl, et al. (2021) presented a model-based approach that jointly considers response and RT information for identifying C/IER. The approach assumes different data-generating processes to underlie responses and RTs associated with attentive and inattentive response behavior. For attentive responses, the model assumes customary IRT models for polytomous data to hold, such as the generalized PCM. For inattentive responses, the model assumes category probabilities to be unrelated to item characteristics and persons' content trait levels and estimates marginal category probabilities of inattentively choosing a given category over all types of C/IER patterns. Note that in doing so the approach avoids assumptions on specific C/IER patterns. Attentive RTs are modeled in line with common RT models for noncognitive data (Ferrando & Lorenzo-Seva, 2007; Molenaar et al., 2015), considering possibly complex relations between respondents' trait levels and RTs. More specifically, the approach considers the distance-difficulty hypothesis, stating that persons who either strongly agree or disagree with a statement can quickly decide on a suitable response option, while persons for whom it is difficult to decide whether or not they agree with a statement need more time for their decision (Kuncel & Fiske, 1974). Inattentive RTs, in turn, are assumed to be unaffected by person and item characteristics and to generally be shorter than attentive RTs.

Attentive Response Styles

In contrast to research on C/IER, where RTs have oftentimes been employed for identifying response bias, in the context of attentive RS, RTs have

predominantly been used for more closely investigating preidentified attentive RS. Henninger and Plieninger (2020), for instance, examined the relationship between acquiescent responding, MRS, and ERS with RTs to draw inferences on cognitive processes in rating scale usage. The authors reported both respondent-level and item-by-respondent level effects of ERS on RTs. They found respondents with higher ERS to require longer time to generate responses, particularly when providing nonextreme responses. The authors interpreted this finding as evidence against the common notion that ERS can be seen as stemming from low cognitive effort of the respondent (referring to, e.g., Aichholzer, 2013; Krosnick, 1999). Instead, Henninger and Plieninger (2020) concluded that respondents with moderate to high ERS seem to give nonextreme responses more deliberately.

Proposed Approach

In questionnaire data, it is likely that different types of response bias occur. Up to now, although models for both RS and C/IER exist, there is no model that incorporates both types of response bias. To jointly account for and investigate C/IER and RS, we propose to leverage the rich information contained in RT data and suggest an RT-based mixture modeling approach. The approach is based on the mixture modeling approach to C/IER by Ulitzsch, Pohl, et al. (2021) and extends it to incorporate RS. To keep the model simple, we identify C/IER on the person level, that is, assume respondents to have a constant probability of showing C/IER across the questionnaire, instead of allowing for attentiveness to vary on the screen-by-respondent level as in Ulitzsch, Pohl, et al. (2021). Class membership of person i ($i = 1, \dots, I$) is denoted with g_i . For attentive respondents, we write $g_i = 1$ and for careless respondents $g_i = 2$. For each class, different component models can be formulated that incorporate researchers' beliefs on how attentive and inattentive respondents interact with noncognitive assessments. To consider attentive RS, a component model for attentive responses can be chosen that accommodates RS. The approach is implemented in a Bayesian framework. In the following, we present the approach in greater detail, making suggestions for all constituting components. The model can be applied to scales measuring a single trait and all items can be coded in the same direction. However, applying it to multiple scales (i.e., using information on multiple traits) and using some reverse-coded items—while preserving the information on chosen response options through negative discrimination parameters—may facilitate estimation (see Ulitzsch et al., 2022). For simplicity, we present the model assuming the same number of response options for all items.

Attentive Behavior

Item responses. Attentive item responses are assumed to be governed by both respondents' content trait levels and their RS. To this end, researchers may employ a model accommodating RS of their choosing. We here present the

approach drawing on the framework for integrating RS into IRT models for polytomous data by Henninger and Meiser (2020), which conceptualizes RS as person-specific shifts in threshold parameters. Other alternatives are outlined in the discussion.

We integrate RS into a multidimensional generalized PCM. For the sake of simplicity, we assume a simple structure for the content traits, that is, each item is assumed to load on one content trait only. We denote respondent i 's response to item j ($j = 1, \dots, J$) with $x_{ij} \in \{0 \dots K\}$, with K giving the highest possible response category on the items considered. In the extended multidimensional generalized PCM accommodating RS according to the framework outlined by Henninger and Meiser (2020), the probability that respondent i chooses category k on the j th item is given by

$$p(x_{ij} = k) = \frac{\exp\left(\sum_{l=0}^k \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + \gamma_{il})\right)\right)}{\sum_{r=0}^K \exp\left(\sum_{l=0}^r \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + \gamma_{il})\right)\right)} \quad (1)$$

with $\sum_{l=0}^0 \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + \gamma_{il})\right) \equiv 0$.

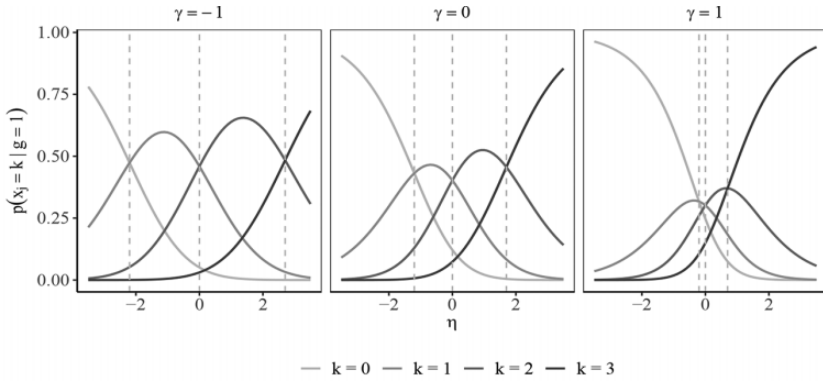
Here, α_j and δ_{jl} give item j 's discrimination and l th threshold parameter. The function $s[j]$ maps the item's index j to the index s ($s = 1, \dots, S$) of the substantive trait η_s it is assumed to measure, and $\eta_{is[j]}$ gives respondent i 's location on that trait. The parameter γ_{il} gives respondent i 's person-specific shift in threshold l .

In the empirical example and the study of parameter recovery, we will exemplarily focus on modeling ERS—one of the most studied theory-derived RS (see Buckley, 2009; Clarke, 2000; Dibek & Cikrikci, 2021; He & Van de Vijver, 2016; Johnson, 2005; Ju & Falk, 2019; Lu & Bolt, 2015, for applications). As noted in Henninger and Meiser (2020) and Henninger (2021), ERS can be accommodated by modeling perfectly negatively correlated shifts in outer threshold parameters (see also Falk & Cai, 2016; Wetzel & Carstensen, 2015). Hence, for modeling ERS, only a single person-specific RS parameter is needed, and the model given in Equation 1 can be simplified to

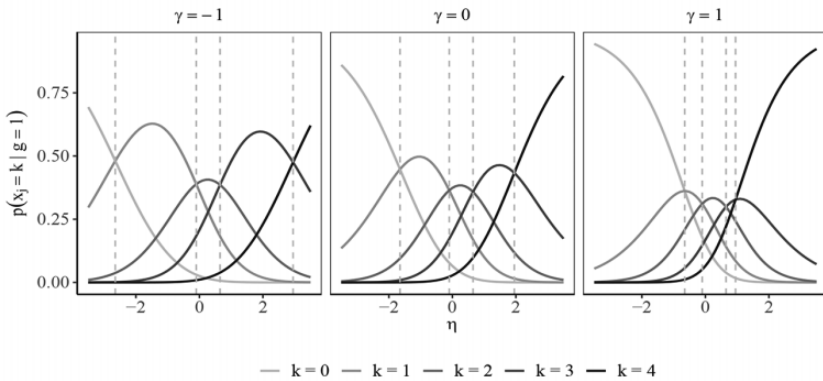
$$p(x_{ij} = k) = \frac{\exp\left(\sum_{l=0}^k \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + c_l \gamma_i)\right)\right)}{\sum_{r=0}^K \exp\left(\sum_{l=0}^r \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + c_l \gamma_i)\right)\right)} \quad (2)$$

with $\sum_{l=0}^0 \alpha_j \left(\eta_{is[j]} - (\delta_{jl} + c_l \gamma_i)\right) \equiv 0$.

Careless Responding and Attentive Response Styles



(a) Four response categories



(b) Five response categories

FIGURE 2. Schematic illustration of person-specific shifts of item thresholds for different locations on the extreme response style trait γ . For the example with five categories, γ is assumed to affect the outer thresholds only. Person-specific thresholds are marked with dashed lines.

In the case of $K = 3$ —which will be considered in the empirical example and the study of parameter recovery—we set $c_1 = 1$, $c_2 = 0$, and $c_3 = -1$ (see Wetzel & Carstensen, 2015, for a different parameterization of the model).¹ Hence, for respondents scoring high on the ERS trait γ , the upper and lower thresholds are coerced toward each other, reflecting respondents' tendency to favor outer (i.e., extreme) over inner (i.e., nonextreme) response categories, while for respondents with negative γ , the upper and lower thresholds are coerced away from each other, resulting in respondents to favor inner over outer response categories. The shifts in item thresholds for different γ are illustrated in Figure 2.

Attentive respondents are assumed to approach all scales with the same level of γ , that is, their ERS tendency is assumed to be equal across items and scales.

Response times. Item-level RTs are denoted with t_{ij} , containing the time respondent i spent on the j th item. Attentive RTs are modeled in line with common approaches for noncognitive assessment data (Ferrando & Lorenzo-Seva, 2007; Molenaar et al., 2015; Ranger, 2013), with log attentive RTs being governed by (a) a time intensity parameter β_j that indicates how much time the item requires to be evaluated and responded to, (b) a person speed parameter τ_i , indicating how fast the respondent generates attentive responses, that is, respondents with higher speed levels require less time to generate attentive responses (see also interpretations of the speed parameter in models for RTs from cognitive assessments in van der Linden, 2007), as well as (c) how strongly the respondent agrees or disagrees with the item content, with respondents being hypothesized to require less time the stronger they agree or disagree. In the literature, this effect is referred to as the distance-difficulty relationship between traits and RTs (Kuncel & Fiske, 1974) and can, among others, be incorporated by regressing log RTs on the absolute weighted distance between the respondent's trait level and the middle threshold parameter o_j (Molenaar et al., 2015).² These considerations lead to modeling attentive RTs as

$$\ln(t_{ij}|g_i = 1) \sim \mathcal{N}(\beta_j - \tau_i - \lambda|\alpha_j(\eta_{is[j]} - o_j)|, \sigma_A^2), \quad (3)$$

with σ_A^2 giving the residual variance of attentive log RTs. The distance-difficulty parameter λ determines the expected reduction in RT due to a large distance between the respondent's trait level and the item's middle threshold parameter. Positive values provide supporting evidence for the distance-difficulty hypothesis. Person speed is assumed to be constant across scales. Note that we consider the distance of the content trait $\eta_{is[j]}$ —with ERS being accounted for—to the middle threshold o_j . As such, the model does not incorporate assumptions on whether or not ERS results in higher RTs. As will be delineated below, differences in pacing between respondents differing in ERS are modeled via the correlation between speed τ and the ERS trait γ .

Careless and Insufficient Effort Behavior

Item responses. When being inattentive and showing C/IER, respondents are assumed to choose response options that do not reflect their trait level, for example, by answering uniformly randomly or marking straight or diagonal lines. Based on these considerations, for inattentive responses, marginal category probabilities κ_k over all types of C/IER patterns are estimated (see Ulitzsch, Pohl, et al., 2021), that is,

$$p(x_{ij} = k | g_i = 2) = \kappa_k \quad \text{with} \quad \sum_{k=0}^K \kappa_k = 1. \quad (4)$$

In a simulation study, Ulitzsch, Pohl, et al. (2021) could show that modeling C/IE responses in terms of marginal category probabilities is well capable of capturing different types of C/IER behavior, ranging from random responding around the mid- or endpoints to structured patterns such as straight and diagonal lining.

Response times. Inattentive RTs are assumed to stem from respondents quickly proceeding through the questionnaire and choosing responses without evaluating the item content. As such, inattentive RTs are assumed to (a) be unaffected by person and item characteristics and (b) to be, on average, shorter than attentive RTs. These assumptions are incorporated into the model by assuming the log-normal distribution of inattentive RTs to be governed by a common mean β_C and a common variance σ_C^2 , both of which are assumed to be the same for all items,

$$\ln(t_{ij} | g_i = 2) \sim \mathcal{N}(\beta_C, \sigma_C^2). \quad (5)$$

Imposing the constraint

$$\beta_j \geq \beta_C \quad (6)$$

on time intensities for attentive RTs β_j ensures that, on average, inattentive RTs are shorter than attentive RTs (see Ulitzsch, Pohl, et al., 2021).

Joint Distribution of Person Parameters

For simplicity, the probability to show attentive response behavior is assumed to be unrelated to trait and speed levels (as in the models for rapid guessing in cognitive assessments by Schnipke & Scrams, 1997; Wang & Xu, 2015). Person parameters of the attentive component models are assumed to be multivariate normally distributed with mean vector and covariance matrix

$$\boldsymbol{\mu} = (\mu_\tau, \mu_\gamma, \mu_{\eta_1}, \dots, \mu_{\eta_S}) \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\tau^2 & \sigma_{\tau\gamma} & \sigma_{\tau\eta_1} & \dots & \sigma_{\tau\eta_S} \\ \sigma_{\tau\gamma} & \sigma_\gamma^2 & \sigma_{\gamma\eta_1} & \dots & \sigma_{\gamma\eta_S} \\ \sigma_{\tau\eta_1} & \sigma_{\gamma\eta_1} & \sigma_{\eta_1}^2 & \dots & \sigma_{\eta_1\eta_S} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{\tau\eta_S} & \sigma_{\gamma\eta_S} & \sigma_{\eta_1\eta_S} & \dots & \sigma_{\eta_S}^2 \end{pmatrix}. \quad (7)$$

For identifying the model, we set person parameter means to zero and content trait variances to one, while leaving discrimination and threshold parameters

unconstrained. Item parameters are modeled as fixed effects. This yields the following likelihood

$$\mathcal{L} = \prod_{i=1}^I \left(\pi_i \prod_{j=1}^J p(x_{ij} | \gamma_i, \eta_{is|j}, \alpha_j, \delta_j)^{(1-d_{ij}^{(x)})} f(t_{ij} | \tau_i, \eta_{is|j}, \beta_j, \lambda, \alpha_j, o_j, \sigma_A^2)^{(1-d_{ij}^{(t)})} + \right. \\ \left. (1 - \pi_i) \prod_{j=1}^J p(x_{ij} | \kappa)^{(1-d_{ij}^{(x)})} f(t_{ij} | \beta_C, \sigma_C^2)^{(1-d_{ij}^{(t)})} \right) h(\boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_S | \boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (8)$$

with $\pi_i = p(g_i = 1)$ giving the probability that respondent i approached the assessment attentively, and $(1 - \pi_i) = p(g_i = 2)$ denoting respondent i 's carelessness probability. The term $h(\boldsymbol{\tau}, \boldsymbol{\gamma}, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_S | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal density of the person parameters of the attentive component models. To consider missing responses and missing RTs, $d_{ij}^{(x)}$ and $d_{ij}^{(t)}$, denoting whether or not a response or RT of respondent i to the j th item is available, are included in the likelihood. We set $d_{ij}^{(x)} = 1$ and $d_{ij}^{(t)} = 1$ for missing responses, respectively, RTs, and set $d_{ij}^{(x)} = 0$, respectively, $d_{ij}^{(t)} = 0$, otherwise. This way, assuming missing at random (MAR), all available data are used. Note that the MAR assumption accommodates different probabilities of occurrence of missing responses under C/IER and attentive behavior.

Prior Distributions

Bayesian estimation techniques facilitate estimation of the proposed approach. Priors are set in accordance with Ulitzsch, Pohl, et al. (2021). For the person parameter covariance matrix, a decomposition strategy is employed (Barnard et al., 2000), with separate prior distributions for correlations and standard deviations. For the correlation matrix, we employ an Lewandowski-Kurowicka-Joe prior (Lewandowski et al., 2009) with shape 1. For model identification, content trait standard deviations are set to unity. For the standard deviations of speed σ_τ and the ERS trait σ_γ , half-Cauchy priors with location 0 and scale 5 are employed. These are also imposed on the residual standard deviation of log attentive RTs σ_A , the common standard deviation of log C/IE RTs σ_C , as well as item discriminations α_j . For thresholds δ_{jl} , the distance-difficulty parameter λ as well as the common mean β_C , we employ diffuse normal priors with mean 0 and standard deviation 10. For time intensity offset parameters $\beta_j^* = \beta_j - \beta_C$, we employ diffuse half-normal priors with mean 0 and standard deviation 10. C/IER category probabilities are equipped with a diffuse Dirichlet prior with $\boldsymbol{\kappa} \sim \text{Dir}(\mathbf{1})$. For attentiveness and C/IER probabilities $(\pi_i, 1 - \pi_i)$, we employ a Dirichlet prior, parameterized as $(\pi_p, 1 - \pi_p) \sim \text{Dir}(v(\pi_p, 1 - \pi_p))$, where $(\pi_p, 1 - \pi_p)$ gives the population-level proportions of attentive and C/IE respondents and v is a concentration parameter (see Kemp et al., 2007; Salakhutdinov

et al., 2012). For population-level proportions of attentive and C/IE respondents ($\pi_p, 1 - \pi_p$), a diffuse Dirichlet prior with $(\pi_p, 1 - \pi_p) \sim \text{Dir}(1, 1)$ is implemented. For the concentration parameter ν , a half-Cauchy prior with location 0 and scale 5 is used.

Empirical Example

The purpose of the empirical example is fourfold. First, we investigate whether in empirical data C/IER and attentive ERS can be distinguished. Second, the empirical example serves to illustrate the insights into response behavior that can be gained on the basis of the proposed model. Third, we assess the impact of neglecting C/IER, ERS, or both. Fourth, we contrast the proposed approach against a two-step approach that first filters out C/IE respondents using threshold-based procedures and then applies an ERS model to the cleaned data set.

Data

We analyzed responses and raw log data from the German subsample ($I = 2,847$) of the PISA 2015 background questionnaire. The PISA 2015 assessment focused on science as the major domain. For illustrating the proposed approach, we focused on environmental awareness and enjoyment of science, measured with 7 and 5 four-point Likert-type scale items, respectively. Items for either scale were presented on a single screen. For measuring environmental awareness, respondents were asked to gauge how informed they are on different environmental issues, for example, nuclear waste or water shortage. Enjoyment of science was measured by asking respondents to express their level of agreement with statements such as “I generally have fun when I am learning science topics”. There were no negatively worded items. A total of 0.07% of item responses were missing. Raw log data were used to reconstruct item-level RTs, making use of the R package logFSM (Kroehne, 2019). The package implements the finite state machine (FSM) framework for log data presented by Kroehne and Goldhammer (2018). In the FSM framework, the RT for an item is reconstructed by taking the difference between the time stamp associated with choosing a response option on that item and the time stamp associated with providing the preceding response. Note that in the FSM framework, the RT for the first item answered cannot be reconstructed as it is confounded with the time taken for reading the question stem. The FSM framework does not require items to be answered in a linear order (see Kroehne et al., 2019, for details).

Analyses

We analyzed the data using the proposed model considering both C/IER and ERS in attentive responding. Further, we specified three special cases of the proposed model, neglecting either C/IER, ERS, or both. In all models, missing

RTs due to the FSM reconstruction were ignored, while the associated responses were considered. Note that doing so entails assuming MAR for such missing RTs, corresponding to the assumption that respondents' decisions on which item to answer first are unrelated to their trait levels, speed, and location on the ERS trait.

In the model considering C/IER but neglecting ERS, attentive item responses were modeled using a generalized PCM, that is, the same thresholds were assumed for all respondents. In the model considering ERS but neglecting C/IER, all responses were assumed to stem from attentive response processes, that is, the mixture component was dropped and all responses and RTs were modeled according to Equations 2 and 3. Finally, in the model considering neither C/IER nor ERS, all item responses were modeled using a generalized PCM and RTs were modeled according to Equation 3. The four models were compared by means of the widely applicable information criterion (WAIC; Vehtari et al., 2017; Watanabe, 2013). We investigated both structural parameters as well as differences in person parameter estimates between the different models.

For implementing a two-step approach to jointly considering C/IER and RS, we first filtered the data for C/IER using a sequential multiple-hurdle procedure (Curran, 2016) that integrates information of multiple C/IER indicators, each being sensitive to a different aspect. In the present analyses, we employed the average time per item, the long string index, and Mahalanobis distance, sequentially filtering out respondents with the most extreme values on these indicators. We then analyzed the filtered data set using a model accommodating ERS, with responses and RTs being modeled according to Equations 2 and 3. Following Ulitzsch, Pohl, et al. (2021), in order to evaluate the range of possible results and the impact of threshold settings, we implemented two sets of thresholds, choosing either a liberal or a conservative cutoff for all three indicators employed. Details on the threshold settings are given in Table 1. For further details on implementation, we refer to Ulitzsch, Pohl, et al. (2021).

All analyses were performed using R Version 3.6.3 (R Development Core Team, 2017). Bayesian estimation was conducted using Stan Version 2.19 (Carpenter et al., 2017) employing the rstan package Version 2.19.3 (Guo et al., 2018). For all models, we ran two MCMC chains with 3,000 iterations each, with the first half being employed as warm-up. Stan code for the most general model accommodating both C/IER and ERS is provided in the OSF repository accompanying this study. The sampling procedure was assessed on the basis of trace plots and potential scale reduction factor (PSRF) values, with PSRF values below 1.10 for all parameters being considered as satisfactory (Gelman & Rubin, 1992; Gelman & Shirley, 2011). WAIC values were computed using the package loo (Vehtari et al., 2020). The long string index and Mahalanobis distance were calculated using the package careless (Yentes & Wilhelm, 2021).

TABLE 1.
Threshold Sets Employed for Identifying Careless and Insufficient Effort Respondents in the Two-Step Approach

	Conservative Threshold Set	Liberal Threshold Set
Average time per item	Below 1 second	Below 2 seconds
Long string index	The same response on all 12 items	The same response on at least five of the seven environmental awareness items and at least four of the five enjoyment of science items
Mahalanobis distance	Exceeding the 95th quantile of a χ^2 distribution with 12 <i>df</i>	Exceeding the 99th quantile of a χ^2 distribution with 12 <i>df</i>

Note. Recall that squared Mahalanobis distance can be approximated by a χ^2 distribution with degrees of freedom corresponding to the number of variables (Rousseeuw & Van Zomeren, 1990).

Results

In all specified models, we observed good mixing of the MCMC chains and no PSRF values above 1.10. WAIC values and person parameter variances and correlations of all specified models are displayed in Table 2. Compared to the models neglecting C/IER, ERS, or both, the proposed model yielded the lowest WAIC, indicating that both response bias were present in the data.

Investigating response behavior. In the proposed model, the population-level proportion of careless respondents (i.e., $1 - \pi_p$) was estimated to be .04 (95% credibility interval: [.03, .04]). Note that this corresponds to the expected rate of careless respondents (i.e., 4%). ERS tendency showed considerable variation across respondents and was weakly negatively related to speed. Thus, attentive respondents with a stronger preference for extreme response options tended to proceed more slowly through the questionnaire. This finding is in line with the respondent-level effect reported by Henninger and Plieninger (2020), who found respondents high in ERS to require more time to generate responses. We did not find respondents with different environmental awareness and enjoyment of science levels to differ in their ERS tendencies, as indicated by correlations not credibly different from zero. With $\lambda = 0.06$ [0.06; 0.07], there was evidence for the distance-difficulty relationship between traits and RTs. Recall that positive values for the distance-difficulty parameter λ result in shorter expected RTs the greater the distance between respondents' trait level and the item's middle threshold parameter. Hence, when the absolute difference between the

TABLE 2.

Person Parameter Variances and Correlations of Different Models of Response Behavior

No Response Bias WAIC: 113,643				ERS Only WAIC: 111,283			
	τ	η_1	η_2	τ	γ	η_1	η_2
τ	0.07 [0.06, 0.07]			0.06 [0.06, 0.07]			
γ				-.08 [-.15, -.02]	0.33 [0.29, 0.37]		
η_1	-.09 [-.13, -.04]	1.00		-.05 [-.10, -.01]	.03 [-.03, .10]	1.00	
η_2	-.04 [-.09, .00]	.44 [.40, .47]	1.00	-.03 [-.08, .00]	.06 [-.01, .12]	.43 [.39, .47]	1.00
C/IER only WAIC: 110,149				C/IER and ERS WAIC: 108,322			
	τ	η_1	η_2	τ	γ	η_1	η_2
	0.07 [0.06, 0.07]			0.06 [0.05, 0.06]			
γ				-.18 [-.19, -.26]	0.29 [0.26, 0.33]		
η_1	-.07 [-.12, -.04]	1.00		-.04 [-.09, .00]	.03 [-.03, .09]	1.00	
η_2	-.04 [-.09, .00]	.43 [.40, .46]	1.00	-.04 [-.08, .01]	.02 [-.05, .08]	.43 [.40, .47]	1.00

Note. 95% Bayesian credibility intervals are given in squared brackets. τ = speed; γ = extreme response style tendency; η_1 = environmental awareness; η_2 = enjoyment of science; C/IER = careless and insufficient effort responding; ERS = extreme response style; WAIC = widely applicable information criterion computed from the item-by-respondent wise log likelihood.

respondent’s trait level and the item’s middle threshold increased by one standard deviation,³ attentive RTs were expected to decrease by the factor $\exp(-0.06) = 0.94$. Further, respondents tended to favor inner (i.e., nonextreme) response options in C/IE responding, as evidenced by a higher κ for inner ($\kappa_1 = .26$ [.22; .30]; $\kappa_2 = .53$ [.48; .58]) as compared to outer (i.e., extreme) response categories ($\kappa_0 = .11$ [.08; .14]; $\kappa_3 = .10$ [.07; .13]). As pointed out in Ulitzsch, Pohl, et al. (2021), this is in line with cognitive theories on edge aversion in decision making processes when items do not need to be (or, as in the present case, are not) processed (Bar-Hillel, 2015).

Investigating the consequences of neglecting response bias. By and large, all models yielded comparable estimates of the correlations between person variables. The models, however, led to somewhat different conclusions concerning

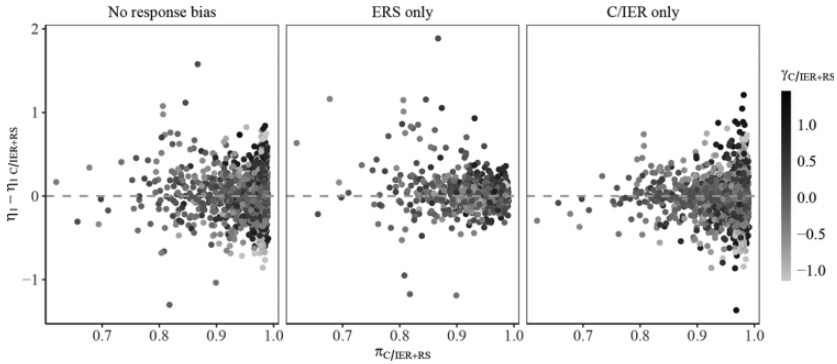


FIGURE 3. Differences in environmental awareness estimates (η_1) of the model considering neither response bias (left panel), the model considering extreme response styles only (middle panel), and the model considering careless and insufficient effort responding only (right panel) to the model considering both types of response bias ($\eta_{1C/IER+RS}$). Differences are plotted against attentiveness probability estimates ($\pi_{C/IER+RS}$) and colored by the extreme response style tendency ($\gamma_{C/IER+RS}$) retrieved from the full model.

response bias. While the model considering only C/IER but neglecting ERS did to not lead to considerable different conclusions on the population-level proportion of careless respondents (.04 [.04, .05]), the model yielded slightly higher marginal C/IER category probabilities for the outer response options ($\kappa_0 = .15$ [.13; .18]; $\kappa_1 = .26$ [.22; .30]; $\kappa_2 = .47$ [.44; .51]; $\kappa_3 = .11$ [.09; .14]). This may indicate that, when neglecting ERS, some respondents with pronounced preferences for the outer categories were deemed more plausible to have shown C/IER. Neglecting C/IER but considering ERS yielded a higher estimate of the variability of respondents' ERS tendencies and a lower correlation between speed and ERS tendency.

While we encountered only small differences in estimates of structural parameters, we observed pronounced differences on the individual (and, as such, possibly subgroup) level. Figure 3 gives differences in environmental awareness estimates (η_1) of the model considering neither response bias (left panel), the model considering ERS only (middle panel), and the model considering C/IER only (right panel) to the model considering both types of response bias ($\eta_{1C/IER+RS}$). Differences are plotted against attentiveness probability estimates ($\pi_{C/IER+RS}$) and colored by ERS tendency ($\gamma_{C/IER+RS}$) retrieved from the full model.⁴ As evidenced in Figure 3, leaving response bias unconsidered may impact the estimates of content trait levels. We observed both upward and downward adjustments of trait estimates when either type of response bias was considered. In the present analyses, when C/IER was not modeled (left and middle panel), differences in estimates of environmental awareness compared to the full

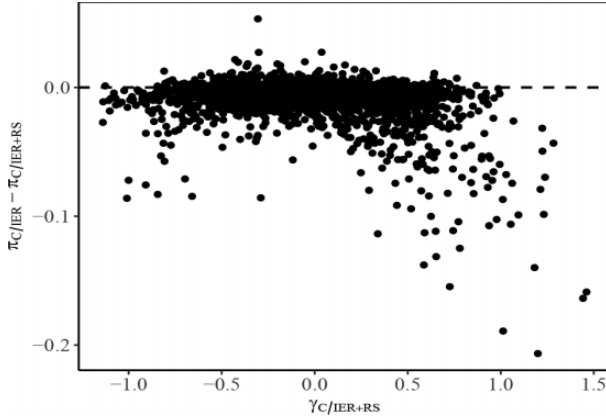


FIGURE 4. Differences in attentiveness probability estimates retrieved from the model considering careless and insufficient effort responding only ($\pi_{C/IER}$) and the model considering both extreme response styles and careless and insufficient effort responding ($\pi_{C/IER+RS}$) plotted against the extreme response style tendency ($\gamma_{C/IER+RS}$) retrieved from the full model.

model increased with decreasing attentiveness probabilities. When leaving ERS unconsidered (left and right panel), differences in environmental awareness were most pronounced for respondents with high attentiveness probabilities who either scored very high (i.e., strongly favored outer over inner response categories in their attentive responding, colored in black) or very low (i.e., strongly favored inner over outer response categories in their attentive responding, colored in light gray) on the ERS trait. We found similar patterns for enjoyment of science trait estimates.

Likewise, conclusions concerning individual response behavior were impacted if only some aspects of response bias were modeled. Figure 4 depicts differences in attentiveness probability estimates retrieved from the model considering C/IER only ($\pi_{C/IER}$) and the full model ($\pi_{C/IER+RS}$). These differences are plotted against ERS parameters retrieved from the full model ($\gamma_{C/IER+RS}$). In the present application, both analyses yielded comparable conclusions concerning attentiveness for respondents located near the mean on the ERS trait ($\gamma \approx 0$), that is, who—relative to other respondents—had no pronounced category preferences. With increasingly extreme locations on the ERS trait (i.e., for large negative and large positive scores on γ), however, the model neglecting ERS yielded lower attentiveness probabilities than the full model. That is, respondents with higher tendencies to favor inner over outer response categories or vice versa were more likely to be deemed inattentive when ERS was neglected. These differences were more strongly pronounced for respondents favoring outer over inner response categories ($\gamma > 0$). This pattern of differences in attentiveness

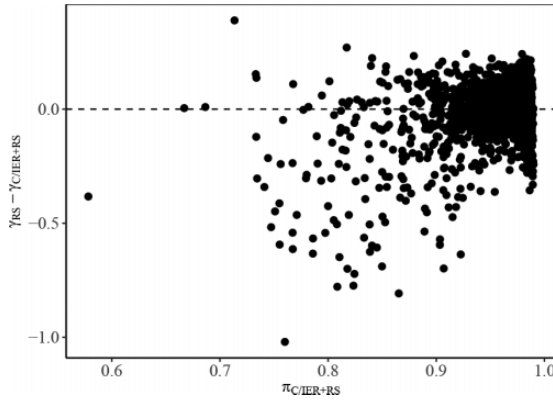


FIGURE 5. Differences in extreme response style tendencies retrieved from the model considering extreme response styles only (γ_{RS}) and the model considering both extreme response styles and careless and insufficient effort responding ($\gamma_{C/IER+RS}$) plotted against attentiveness probabilities ($\pi_{C/IER+RS}$) retrieved from the model considering both response biases.

probabilities may trace back to the fact that respondents with extreme locations on the ERS trait tended to generate responses that do not well align with the attentive component model neglecting ERS, and, therefore, tended to get “absorbed” by the inattentive component model (which incorporates less restrictive assumptions on category probabilities).

Figure 5 depicts differences in ERS parameter estimates retrieved from the model considering ERS only (γ_{RS}) and the full model ($\gamma_{C/IER+RS}$) plotted against attentiveness probabilities retrieved from the full model ($\pi_{C/IER+RS}$). Both analyses yielded comparable conclusions concerning ERS for respondents with high attentiveness probabilities. We assume that the observed differences in ERS for respondents with high attentiveness probabilities go back to differences in item parameter estimates between the models. With decreasing attentiveness probabilities, the model neglecting C/IER yielded lower estimates for γ than the full model, indicating a higher tendency to favor inner over outer response categories. Recall that in the present application, C/IE respondents tended to favor inner over outer response categories. That is, (presumed) C/IE behavior tended to get “absorbed” by the ERS trait.

In sum, Figures 3 through 5 illustrate size and nature of the differences in person parameter estimates that can be encountered in empirical settings when different types of response bias are accounted for. These differences in person parameter estimates pose considerable effects and indicate that, in the case that a subgroup variable relates to response bias, modeling versus neglecting response bias may also impact subgroup results.

TABLE 3.

Person Parameter Variances and Correlations of the Model Accommodating Extreme Response Styles After Filtering for Careless Respondents

	Conservative Threshold Set				Liberal Threshold Set			
	τ	γ	η_1	η_2	τ	γ	η_1	
τ	0.04				0.02			
	[0.04, 0.04]				[0.02, 0.02]			
γ	-.18	0.18			-.26	0.14		
	[-.25, -.10]	[0.15, 0.21]			[-.38, -.14]	[0.11, 0.17]		
η_1	-.03	.04	1.00		.11	.13	1.00	
	[-.08, .02]	[-.04, .13]			[.03, .18]	[.04, .22]		
η_2	-.12	-.01	.39	1.00	-.05	-.02	.45	1.00
	[-.17, -.07]	[-.10, .07]	[.35, .42]		[-.12, .01]	[-.11, .09]	[.42, .49]	

Note. 95% Bayesian credibility intervals are given in squared brackets. τ = speed; γ = extreme response style tendency; η_1 = environmental awareness; η_2 = enjoyment of science; C/IER = careless and insufficient effort responding; ERS = extreme response styles.

Comparisons with a two-step approach to accounting for multiple response bias. The conservative and liberal threshold settings filtered out 9.91% and 22.83% of respondents, respectively, both by far exceeding the 4% implied by the proposed model-based approach. Table 3 displays person parameter variances and correlations of the ERS model applied to the filtered data sets. Note that the models cannot be compared in terms of the WAIC, as the filtered data sets comprise different respondents. Results for the implementations of the two-step approach considerably differed from those displayed in Table 3. ERS variances, for instance, were much lower in the applications of the two-step approach, presumably due to classifying more respondents with identical responses to all (conservative threshold set) or the majority of the items (liberal threshold set) as careless, while some of these respondents were deemed more plausible to have displayed ERS in the model-based approach. More importantly, results for the two implementations of the two-step approach vastly differed from each other, further illustrating that results of filtering-based methods are heavily dependent on threshold settings, with vast differences being observable even for small differences in the employed thresholds.

Parameter Recovery

To investigate the trustworthiness of results obtained in the empirical application, we conducted a parameter recovery study, with data-generating

parameters mimicking those obtained from the application of the proposed model to PISA data.

Data Generation

We generated 100 data sets for $I = 500$ respondents being administered items from two scales, each comprising four items with a four-point Likert-type scale. We considered a population-level proportion of C/IE respondents of .05. Standard deviations of content traits and their correlation were set to $\sigma_{\eta_1} = \sigma_{\eta_2} = 1.00$ and $cor(\eta_1, \eta_2) = .40$. Correlations of the content traits with speed and the ERS trait were set to $cor(\tau, \eta_1) = cor(\tau, \eta_2) = cor(\gamma, \eta_1) = cor(\gamma, \eta_2) = 0$. Standard deviations of speed and the ERS trait were set to $\sigma_\tau = \sqrt{0.04} = 0.20$ and $\sigma_\gamma = \sqrt{0.30} = 0.55$. The correlation between speed and the ERS trait was set to $cor(\tau, \gamma) = -.20$. Item discriminations α_j were set to 1. For each item, thresholds were simulated to be equally spaced in steps of 0.50, with the middle threshold δ_{j2} drawn from the set $\{-1.00, -0.50, 0.50, 1.00\}$. Time intensities β_j were drawn from the set $\{0.95, 1.20, 1.20, 1.45\}$. The distance-difficulty parameter was set to $\lambda = 0.05$. Residual standard deviations of attentive log RTs and the common standard deviation of inattentive RTs were set to $\sigma_A = 0.50$ and $\sigma_C = 1.25$. The common mean of log inattentive RTs was set to $\mu_C = 0.70$.

Following Curran and Denison (2019) and Ulitzsch, Pohl, et al. (2021), we considered a scenario with different C/IER patterns, thereby illustrating that the model can handle the joint occurrence of different C/IER patterns. Inattentive respondents were randomly partitioned into three equally sized groups, representing uniform random responding, straight lining, and diagonal lining. For each group, patterns were generated to result in equal marginal response categories for all categories, such that marginal probabilities for C/IE responses across all patterns were given by $\kappa = (.25, .25, .25, .25)$. To simulate uniform random responding, responses were randomly drawn from a discrete uniform distribution. For generating straight-lining behavior, the first C/IE response was chosen randomly with equal category probabilities and all subsequent responses were set to be the same as the first. For simulating diagonal lining, the first answer was determined randomly to either correspond to the most upper or most lower response category. Responses to the remaining items were then simulated as diagonal lines moving away from the first answer by one category per item.

Estimation

For estimation, we employed the same setup as in the empirical application. To avoid nonconvergence due to an insufficient number of iterations, we used 25,000 iterations for each of the two MCMC chains, with the first half being employed as warm-up.

TABLE 4.
Simulation Results for Selected Parameters

Parameter	True	Median EAP
$1 - \pi_p$	0.05	0.06 [0.05, 0.07]
σ_τ	0.20	0.20 [0.20, 0.21]
σ_γ	0.55	0.56 [0.52, 0.60]
$cor(\tau, \gamma)$	-0.20	-0.20 [-0.25, -0.14]
$cor(\tau, \eta_1)$	0.00	0.00 [-0.03, 0.04]
$cor(\gamma, \eta_1)$	0.00	0.00 [-0.07, 0.07]
$cor(\eta_1, \eta_2)$	0.40	0.40 [0.38, 0.44]
λ	0.05	0.05 [0.04, 0.06]
κ_0	0.25	0.24 [0.20, 0.27]
α_1	1.00	1.05 [0.95, 1.16]
σ_A	0.50	0.49 [0.49, 0.50]
μ_C	0.70	0.70 [0.61, 0.77]
σ_C	1.25	1.24 [1.17, 1.29]

Note. Squared brackets give interquartile ranges of parameter estimates across replications. $1 - \pi_p$ = population-level proportion of careless respondents; τ = speed; γ = extreme response style tendency; η = content trait; λ = distance-difficulty parameter; μ_C and σ_C = common mean and standard deviation of log careless and insufficient effort response times; σ_A = residual standard deviation of log attentive response times; κ_0 = marginal category probability for the first response option; α_1 = item discrimination for the first item; EAP = expected a posteriori.

Results

By and large, we observed good mixing of the MCMC chains. PSRF values above 1.10 were encountered in five of the 100 replications.⁵ For investigating parameter recovery, we considered only replications with all PSRF values below 1.10.

With a median correlation between true and estimated parameters of .99 and .98, item thresholds and time intensity offsets were well recovered. Table 4 displays median EAPs and interquartile ranges of the population-level proportion of careless respondents, person parameter standard deviations and correlations, the distance-difficulty parameter, the common mean and standard deviation of log C/IE RTs, the residual standard deviation of log attentive RTs, and—exemplary—the marginal C/IER category probability for the first response option as well as the item discrimination for the first item alongside the data-generating values. These parameters were estimated without bias (i.e., median EAPs were very close to the true parameters) and exhibited low variability (i.e., were precise, as indicated by narrow interquartile ranges), even with the relatively small considered sample size of $I = 500$, illustrating that the model yields trustworthy parameter estimates under realistic research conditions.

Discussion

We presented a flexible RT-based mixture modeling approach that supports jointly considering, distinguishing, and studying *C/IER* and attentive *RS* in non-cognitive assessment data. *C/IER* and attentive *RS* pose two commonly encountered response bias in questionnaire data. While the former results in responses that are completely uninformative of the traits of interest, the latter results in responses that, although containing information on respondents' levels on the content traits, are confounded with content-irrelevant variability due to differences in category usage. Distinguishing and jointly considering these two types of behavior assists in drawing more valid conclusions from questionnaire data as well as getting a more nuanced understanding of response behavior. The approach has been illustrated on large-scale assessment background questionnaire data but is applicable to any type of computerized questionnaire for which RT data are available (e.g., online surveys).

For separating attentive from inattentive responding, the approach utilizes item-level RT data. In the presented model, RTs serve a twofold purpose. First, considering this rich source of information on response behavior supports separating different types of behavior that often result in rather similar response vectors. Second, in more general terms, considering RT information in models for noncognitive assessment data may enrich the understanding of how respondents interact with such assessments, for example, by investigating whether respondents with different trait levels differ in how fast they generate attentive responses, whether attentive *RS* are related to pacing behavior, or whether there is evidence for the distance-difficulty hypothesis in empirical data.

We applied the proposed model to empirical data from the PISA 2015 background questionnaire, where we found evidence for the joint occurrence of both *ERS* in attentive responding and *C/IER*. The empirical example highlights the potential of the proposed model for understanding processes underlying responses in questionnaire data. To investigate different response bias and get an understanding of their occurrence, the full model that considers both attentive *RS* and *C/IER* is necessary.

In the present application, we found that neglecting either type of response bias may impact conclusions concerning respondents' content trait levels. Further, when either *ERS* or *C/IER* was left unconsidered, the modeled response bias in parts "absorbed" the unconsidered response bias, and respondents with more extreme locations on the *ERS* trait were estimated to have higher carelessness probabilities when *C/IER*, but not *ERS* was modeled, and vice versa. From a conceptual point of view, such effects seem plausible, as different response bias may result in very similar response vectors. Although we observed that the different models yielded different conclusions, it should also be noted that based on the empirical example alone it cannot yet be concluded that neglecting either type of response bias generally yields biased person or subgroup parameter estimates.

We further contrasted the proposed fully model-based approach against two-step approaches to jointly considering C/IER and attentive RS where, in step one, C/IE respondents are filtered out by means of threshold-based procedures and in step two, an IRT model accommodating attentive RS is applied to the cleaned data set. We see two major advantages of the proposed mixture modeling approach over two-step approaches. First, the proposed mixture modeling approach does not rely on threshold settings. There are no globally applicable values for these thresholds, as the distributions of indicators for careless and attentive respondents are scale-specific (Curran, 2016), such that threshold settings are always somewhat arbitrary. Second, the proposed approach differentiates between different types of bias in a single step and thereby avoids the sequential decision procedure of two-step approaches. One advantage of doing so is that the uncertainty of classification is taken into account. This may, for instance, be of relevance when responses and RTs of some C/IE respondents and respondents with certain types of attentive RS are very similar. While two-step procedures require a clear-cut decision for such cases, the proposed approach takes the uncertainty of classification into account. These advantages were illustrated in the empirical example, where we found large differences in structural parameter estimates for small differences in threshold settings. In fact, differences between different implementations of the two-step approach were much more pronounced than differences between fully model-based approaches accommodating different types of response bias. The price for these advantages, however, is increased model complexity, that may result in long running times with increasing questionnaire length and sample size. When these become impractical, heuristic indicator-based approaches may still be the better option for gauging the extent of C/IER in the data at hand.

The approach offers researchers a high degree of flexibility in that different component models can be plugged in for attentive and inattentive responses and RTs, thereby allowing to incorporate specific hypotheses on response behavior as well as to distinguish and study different types of response bias. Researchers may also determine the type of RS component model to employ by means of model comparisons between models with competing component models. For instance, we applied the approach using a model accommodating ERS, derived from the framework presented by Henninger and Meiser (2020). For scales that include midpoint response options, the model can be extended to jointly accommodate ERS and MRS in attentive responding (as in Wetzel & Carstensen, 2015). If researchers have deviating hypotheses concerning the nature of attentive RS that may be present in the data, other component models can be chosen. For deciding on a component model, readers are referred to the unifying framework, overview, and guidelines provided by Henninger and Meiser (2020). It should, nevertheless, be noted that some component models may result in a model that is more challenging to estimate. Mixture models for RS (e.g., Rost, 1991), for instance, would result in a mixture of mixtures. Note that the approach is not limited to PCM or

RSM extensions for modeling attentive RS but may also be extended to IRTree approaches using binary pseudo items. This may be achieved by employing the mixture IRT approach for item responses with different structures by Tijmstra et al. (2018), assuming an IRTree structure based on binary pseudo items for attentive responses and estimating C/IER category probabilities based on polytomous item responses.

Concerning the modeling of attentive RTs, we point out that different approaches exist to incorporating the distance-difficulty relationship between traits and RTs (see Ranger, 2013, for an overview and comparison). Further, the relationship between the distance between the respondent's trait level and the middle threshold parameter and RTs must not necessarily be linear but may take other functional forms (Molenaar et al., 2021). Another alternative to consider may be to use the distance from the item location rather than from the middle threshold parameter. If the distance-difficulty parameter is of substantive interest to researchers, different specifications may be compared by means of model comparisons. Further, the approach allows for incorporating other component models for attentive RTs that support greater flexibility in the assumed RT distribution, e.g., models being based on the Box–Cox normal distribution (Entink et al., 2009) or on categorized RTs (Molenaar et al., 2018).

We found the model to yield good parameter recovery with a sample size of 500, two scales with four items each, and an overall carelessness rate of .05. From these results and results of previous, similar models, we expect that convergence and parameter recovery is good in applications that have comparable or larger sample sizes and carelessness rates. Note that we cannot evaluate all possible parameter constellations and model specifications. Thus, we point out that the statistical performance of the proposed approach may not generalize to all possible combinations of component models. Especially when choosing more complex component models, we advise investigating parameter recovery of the chosen combination of component models to corroborate plausibility of results. The code for our simulation study, which may be adapted to other model specifications, can be found in the OSF repository accompanying this article.

Limitations and Future Research

We found the proposed approach to show good parameter recovery under realistic research conditions. Nevertheless, establishing boundary conditions for which the presented approach may well separate C/IER from attentive RS remains an open and important research question. Challenging conditions that may threaten parameter recovery or the trustworthiness of model comparisons may, for instance, arise under too similar RT distributions of C/IER and attentive responding.

In the empirical application, we found small differences in conclusions on structural parameters drawn from the proposed approach, considering both ERS

and C/IER, and approaches neglecting either ERS, C/IER, or both. Further studies are needed to investigate whether and under which conditions neglecting specific response bias may impact conclusions more heavily. This may, for instance, be the case under higher prevalence of C/IER (recall that in the empirical example, the prevalence was only 4%), when respondents predominantly show C/IER behaviors that result in response vectors resembling those encountered under ERS (i.e., predominantly straight line), or when the ERS trait is more strongly correlated with the content traits. Further, it remains to be investigated under which conditions it is sufficient to account for either C/IER or RS in attentive responding when response bias themselves are not of substantive interest but the objective of analysis is to merely account for response bias.

The presented approach assumes the probability that a respondent provides a C/IE response to be constant across all items considered. Note that this assumption is in line with classical indicator-based procedures drawing on response-pattern-based indicators that also filter at the respondent level. C/IER, however, may vary across the questionnaire and respondents who display C/IER on some parts of the questionnaire might still provide valid responses to others, especially across lengthy questionnaires (Bowling et al., 2020; Gibson & Bowling, 2019). This issue can be accommodated by modeling C/IER behavior on the item-by-respondent (Ulitzsch et al., 2020, 2022) or screen-by-respondent level (Ulitzsch, Pohl, et al., 2021) taking a latent response approach. While integrating such extensions with the proposed model is straightforward, this would result in a highly complex model that is challenging to estimate.

In the context of questionnaires, item-level RT data become increasingly available (see Henninger & Plieninger, 2020, for recent studies recording item-level RTs) or can be reconstructed using the FSM approach (Kroehne, 2019; Kroehne & Goldhammer, 2018). Nevertheless, item-level RTs may not always be at hand. Hence, model adaptations drawing on screen-level RTs (as in Ulitzsch, Pohl, et al., 2021), which can more easily be recorded, or models relying on responses only (as in Ulitzsch et al., 2022) pose a further important topic for future research.

The presented approach showcased the utility of RTs for better understanding how respondents interact with questionnaires. Depending on how questionnaires are administered, researchers may consider additional data such as switches in browser tabs (see Steger et al., 2020, for an application) for identifying respondents not sufficiently engaged with the questionnaire. For instance, the proposed approach may be extended by incorporating the assumption that inattentive respondents may frequently switch to other browser tabs, getting distracted from the questionnaire, while attentive respondents do commonly not display such behavior. Such additional information may also be of great utility for better separating different types of response bias.

Note that the presented approach can entail rather long run times—the model in the empirical application, for instance, required approximately 8 hours to run.

A potential remedy may be the development of maximum likelihood implementations, posing an important topic for future research (see Nagy & Ulitzsch, 2021; Nagy et al., 2022, for implementations of neighboring models for rapid guessing).

Finally, we point out that validation studies are urgently needed for ensuring that the substantive interpretations of the model parameters hold true. Validity could be investigated with experimental manipulations, for example, by varying instructions (as in Bowling et al., 2020; Niessen et al., 2016), through investigations of the model's capability to detect differences between groups of respondents that can be assumed to differ in their levels of C/IER and/or their stylistic tendencies in attentive responding (see Ulitzsch, Penk, et al., 2021, for a validation study using such group comparisons to gain validity evidence for a model-based approach to rapid guessing behavior), or by investigating how attentive RS and adjusted content traits relate to external variables, assuming that relationships adjusted for response bias should more strongly align with subject-matter theory than their unadjusted counterparts (Khorramdel et al., 2017), and that attentive RS and content traits should be linked selectively to extraneous criteria of attentive RS and content traits (Plieninger & Meiser, 2014).

Authors' Note

Supplemental online materials for this article can be found in the OSF and are available via the following link: <https://osf.io/smpn3/>.


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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Esther Ulitzsch  <https://orcid.org/0000-0002-9267-8542>

Steffi Pohl  <https://orcid.org/0000-0002-5178-8171>

Notes

1. For $K = 4$, researchers may set $c_1 = 1$, $c_2 = c_3 = 0$, and $c_4 = -1$ for incorporating the assumption that extreme response styles affect the outer thresholds only, or, for example, $c_1 = 1$, $c_2 = c_3 = \iota$, and $c_4 = -1$, with ι being freely estimated, when they assume that both inner and outer thresholds are affected and draw category preferences away from the midpoint.

2. For instance, for four response categories with three thresholds δ_{j1} , δ_{j2} , and δ_{j3} , o_j corresponds to δ_{j2} . For an uneven number of answer categories, Molenaar et al. (2015) suggested to take the average of the two middle thresholds.
3. Recall that trait variances were set to one for model identification.
4. In the present example, the lowest attentiveness probability was still as high as .62, indicating that none of the respondents was deemed to be careless with high certainty. Rather, some respondents tended to show response and RT patterns that were ambiguous as to whether they represented careless or attentive responding, resulting in their data being downweighed according to their attentiveness probabilities in the estimation of content traits and speed.
5. Note that in practice, researchers would increase the number of iterations even further until the criterion of all potential scale reduction factor values being below 1.10 is fulfilled. We did not explore how many iterations would be sufficient in the present context simply due to computational constraints.

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>
- Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 31(1), 84–98. <https://doi.org/10.1111/j.2044-8317.1978.tb00575.x>
- Bar-Hillel, M. (2015). Position effects in choice from simultaneous displays: A conundrum solved. *Perspectives on Psychological Science*, 10(4), 419–433.
- Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4), 1281–1311.
- 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>
- Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, 24, 1–21. <https://doi.org/10.1177/1094428120947794>
- Buckley, J. (2009). Cross-national response styles in international educational assessments: Evidence from PISA 2006. *NCES Conference on the Program for International Student Assessment: What We Can Learn From PISA*. https://edsurveys.rti.org/pisa/documents/buckley_pisaresponsestyle.pdf
- 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>
- Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, 15(1), 137–152.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19.
- Curran, P. G., & Denison, A. J. (2019). Creating carelessness: A comparative analysis of common techniques for the simulation of careless responder data. <https://psyarxiv.com/ge6fa/>
- Dibek, M. I., & Cikrikci, R. N. (2021). Modelling of the attitude-achievement paradox in TIMSS 2015 with respect to the extreme response style using multidimensional item response theory. *International Journal of Progressive Education*, 17(2), 194–209.
- Entink, R. K., van der Linden, W., & Fox, J.-P. (2009). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, 62(3), 621–640. <https://doi.org/10.1348/000711008X374126>
- 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>
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472. <https://doi.org/10.1214/ss/1177011136>
- Gelman, A., & Shirley, K. (2011). Inference from simulations and monitoring convergence. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (pp. 163–174). Chapman Hall.
- Gibson, A. M., & Bowling, N. A. (2019). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000526>
- Guo, J., Gabry, J., & Goodrich, B. (2018). *Rstan: R interface to Stan* [R package version 2.18.2]. <https://CRAN.R-project.org/package=rstan>
- He, J., & Van de Vijver, F. (2016). Correcting for scale usage differences among Latin American countries, Portugal, and Spain in PISA. *Electronic Journal of Educational Research, Assessment and Evaluation*, 22(1). <https://doi.org/10.7203/relieve.22.1.8282>
- Henninger, M. (2021). A novel partial credit extension using varying thresholds to account for response tendencies. *Journal of Educational Measurement*, 58(1), 104–129. <https://doi.org/10.1111/jedm.12268>
- 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, 1–19. <https://doi.org/10.1177/1073191119900003>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27(1), 99–114.

- 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>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129.
- Ju, U., & Falk, C. F. (2019). Modeling response styles in cross-country self-reports: An application of a multilevel multidimensional nominal response model. *Journal of Educational Measurement*, *56*(1), 169–191. <https://doi.org/10.1111/jedm.12205>
- Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321. <https://doi.org/10.1111/j.1467-7687.2007.00585.x>
- Khorramdel, L., von Davier, M., Bertling, J. P., Roberts, R. D., & Kyllonen, P. C. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study. *Psychological Test and Assessment Modeling*, *59*(1), 71.
- 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. (2020). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*, 0013164420913915. <https://doi.org/10.1177/0013164420913915>
- Kroehne, U. (2019). *LogFSM: Analysis of log data using finite-state machines*. <http://www.logfsm.com/>
- Kroehne, U., Buchholz, J., & Goldhammer, F. (2019, April). *Detecting carelessly invalid responses in item sets using item-level response times (tech. rep.)*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. Toronto, Canada.
- Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563.
- Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *1*, 537–567.
- Kuncel, R. B., & Fiske, D. W. (1974). Stability of response process and response. *Educational and Psychological Measurement*, *34*(4), 743–755. <https://doi.org/10.1177/00131644740.3400401>
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. <https://doi.org/10.1016/j.jmva.2009.04.008>
- Lu, Y., & Bolt, D. M. (2015). Examining the attitude-achievement paradox in PISA using a multilevel multidimensional IRT model for extreme response style. *Large-Scale Assessments in Education*, *3*(1), 1–18. <https://doi.org/10.1186/s40536-015-0012-0>
- Lyu, W., & Bolt, D. M. (2022). A psychometric model for respondent-level anchoring on self-report rating scale instruments. *British Journal of Mathematical and Statistical Psychology*, *75*(1), 116–135. <https://doi.org/10.1111/bmsp.12251>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437. <https://doi.org/10.1037/a0028085>
- 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>
- Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 205–228. <https://doi.org/10.1111/bmsp.12117>
- Molenaar, D., Rózsa, S., & Kő, N. (2021). Modeling asymmetry in the time–distance relation of ordinal personality items. *Applied Psychological Measurement, 45*(3), 178–194. <https://doi.org/10.1177/0146621621990756>
- Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*(1), 56–74. <https://doi.org/10.1080/00273171.2014.962684>
- Nagy, G., & Ulitzsch, E. (2021). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement, 82*. <https://doi.org/10.1177/00131644211045351>
- Nagy, G., Ulitzsch, E., & Lindner, M. A. (2022). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*. <https://doi.org/10.1111/jcal.12719>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality, 63*, 1–11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- Organization for Economic Cooperation and Development. (2017). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf
- Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement, 74*(5), 875–899.
- R Development Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>
- Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling, 55*(4), 361–382.
- Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*(1), 75–92. <https://doi.org/10.1111/j.2044-8317.1991.tb00951.x>
- Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633–639.
- Salakhutdinov, R., Tenenbaum, J. B., & Torralba, A. (2012). Learning with hierarchical-deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1958–1971.

- Samejima, F. (2016). Graded response models. In *Handbook of item response theory* (pp. 123–136). Chapman; Hall/CRC.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. <https://doi.org/10.1111/j.1745-3984.1997.tb00516.x>
- Schroeders, U., Schmidt, C., & Gnams, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement, 82*(1), 29–56. <https://doi.org/10.1177/00131644211004708>
- Steger, D., Schroeders, U., & Wilhelm, O. (2020). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment, 19*(1), 1120914970. <https://doi.org/10.1177/1073191120914970>
- 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>
- Tutz, G., Schauberger, G., & Berger, M. (2018). Response styles in the partial credit model. *Applied Psychological Measurement, 42*(6), 407–427. <https://doi.org/10.1177/0146621617748322>
- Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment, 26*(2), 104–124. <https://doi.org/10.1080/10627197.2020.1858786>
- Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2021). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 86*(1), 11336–021-09817-7. <https://doi.org/10.1007/s11336-021-09817-7>
- Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2023). Accounting for careless and insufficient effort responding in large-scale survey data—Development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods, 55*(1), 022-02053-6. <https://doi.org/10.3758/s13428-022-02053-6>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level nonresponse. *British Journal of Mathematical and Statistical Psychology, 93*(1), 11111/bmsp.12188. <https://doi.org/10.1111/bmsp.12188>
- Ulitzsch, E., Yildirim-Erbaşlı, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in survey data. *British Journal of Mathematical and Statistical Psychology, 93*(1), 11111/bmsp.12272. <https://doi.org/10.1111/bmsp.12272>
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). Loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models [R package version 2.4.1]. <https://mc-stan.org/loo/>
- 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>

- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. <https://doi.org/10.1111/bmsp.12054>
- Wang, W.-C., & Wu, S.-L. (2011). The random-effect generalized rating scale model. *Journal of Educational Measurement*, *48*(4), 441–456. <https://doi.org/10.1111/j.1745-3984.2011.00154.x>
- Watanabe, S. (2013). A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, *14*, 867–897.
- Wetzel, E., & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*. <https://doi.org/10.1027/1015-5759/a000291>
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. <https://doi.org/10.1111/emip.12165>
- Yentes, R. D., & Wilhelm, F. (2021). Careless: Procedures for computing indices of careless responding [R package version 1.2.1].

Authors

ESTHER ULITZSCH is a research scientist at IPN—Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, Kiel 24118, Germany; e-mail: ulitzsch@ipn.uni-kiel.de. Her research interests are psychometrics, process data analysis for investigating response behavior, and Bayesian modeling.

STEFFI POHL is a full professor at Freie Universität Berlin, Habelschwerdter Allee 45, Berlin 14195, Germany; e-mail: steffi.pohl@fu-berlin.de. Her research interests are psychometrics, response behavior, and causal inference.

LALE KHORRAMDEL is the Research Director for Digital Assessment Development and Special Projects at the TIMSS & PIRLS International Study Center at Boston College, 194 Beacon Street, Chestnut Hill, MA 02467, USA; e-mail: lale.khorramdel@bc.edu. Her research interests are psychometrics, response behavior, and digital assessment innovations.

ULF KROEHNE is a post-doc researcher at DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt am Main, Germany; e-mail: u.kroehne@dipf.de. His research interests include psychometrics, technology-based assessment, and log and process data.

MATTHIAS VON DAVIER is the J. Donald Monan, S.J., University Professor in Education at the Lynch School of Education at Boston College as well as the Executive Director of the TIMSS & PIRLS International Study Center at Boston College, 194 Beacon Street, Chestnut Hill, MA 02467, USA; e-mail: matthias.vondavier@bc.edu. His research interests are psychometrics, computational statistics, large-scale international assessments, and artificial intelligence use in assessment.

Manuscript received December 6, 2021

Revision received January 15, 2023

Accepted April 4, 2023