MULTIVARIATE BEHAVIORAL RESEARCH
2024, VOL. 59, NO. 5, 1043–1057

OPEN ACCESS

The Effects of Questionnaire Length on the Relative Impact of Response Styles in Ambulatory Assessment

Kilian Hasselhorn[a], Charlotte Ottenstein[a], Thorsten Meiser[b], and Tanja Lischetzke[a]

[a]Department of Psychology, RPTU Kaiserslautern-Landau, Landau, Germany; [b]University of Mannheim, Mannheim, Germany

ABSTRACT

Ambulatory assessment (AA) is becoming an increasingly popular research method in the fields of psychology and life science. Nevertheless, knowledge about the effects that design choices, such as questionnaire length (i.e., number of items per questionnaire), have on AA data quality is still surprisingly restricted. Additionally, response styles (RS), which threaten data quality, have hardly been analyzed in the context of AA. The aim of the current research was to experimentally manipulate questionnaire length and investigate the association between questionnaire length and RS in an AA study. We expected that the group with the longer (82-item) questionnaire would show greater reliance on RS relative to the substantive traits than the group with the shorter (33-item) questionnaire. Students ($n = 284$) received questionnaires three times a day for 14 days. We used a multigroup two-dimensional item response tree model in a multilevel structural equation modeling framework to estimate midpoint and extreme RS in our AA study. We found that the long questionnaire group showed a greater reliance on RS relative to trait-based processes than the short questionnaire group. Although further validation of our findings is necessary, we hope that researchers consider our findings when planning an AA study in the future.

KEYWORDS

ambulatory assessment; questionnaire length; response styles; multilevel structural equation modeling; multidimensional IRTree models

Ambulatory assessment (AA) is becoming an increasingly popular research method in the fields of psychology and life science (Hamaker & Wichers, 2017). AA (which is an umbrella term for daily diary, experience sampling, and ecological momentary assessment) can be used to assess daily life experiences, such as ongoing behaviors, experiences, physiology, and environmental aspects of people in naturalistic and unconstrained settings (Bolger & Laurenceau, 2013; Fahrenberg, 2006). By applying AA, researchers can study within-person dynamics (e.g., within-person relationships between time-varying variables) in addition to individual differences in these within-person dynamics (Hamaker & Wichers, 2017). Furthermore, the application of AA can reduce recall bias and increase ecological validity (Mehl & Conner, 2012; Trull & Ebner-Priemer, 2014).

When designing an AA study, researchers must make decisions about multiple design features in order to strike a balance between being able to gather rich information, not compromising aspects of AA data (e.g., data quantity and data quality; Arslan et al., 2020; May et al., 2018), and ensuring that they do not overburden their participants (Carpenter et al., 2016). These decisions involve the types of reports to include (e.g., time-based, event-triggered), the number of days to survey people, the number of assessments to administer per day (sampling frequency), and the number of items to administer per measurement occasion (questionnaire length). Mehl & Conner (2012) provide a detailed explanation about study design considerations and methods of data collection in AA.

Recently, researchers have begun to use experimental designs in AA studies to examine the effects of design choices on data quantity and data quality. As outcome variables, experimental AA studies have primarily focused on participant burden, compliance, mean levels of constructs of interest, within-person

variability, within-person relationships, and careless responding (Eisele et al., 2022; Hasselhorn et al., 2022, 2023; Himmelstein et al., 2019). To our knowledge, no study to date has investigated the effect of AA design features on the use of response styles (RS) as a reflection of heuristic processing. Therefore, the aim of the present research was to model RS in AA data, which is nested in structure (measurement occasions nested in individuals), by applying a multilevel model of Item Response Theory (IRT) capturing RS-based and trait-based responses, and to investigate whether questionnaire length influences the (relative) impact of RS on responses in an AA study. We chose to focus on questionnaire length as the design feature of interest because there is some empirical evidence that asking participants to answer more (vs. fewer) items per measurement occasion may lead to reduced data quality (see next section). In the remainder of the Introduction, we first briefly summarize previous results on the effect of questionnaire length on aspects of data quality in AA. Second, we define RS and address the relevance of modeling and accounting for RS in psychological measurement. Third, we describe how RS can be modeled (and accounted for) by current IRT approaches.

The effect of questionnaire length on aspects of data quality in AA

In the experimental AA study by Eisele et al. (2022), longer questionnaires (60 items per measurement occasion) were associated with higher levels of retrospective self-reported careless responding than shorter questionnaires (30 items per measurement occasion). In our own previous research (Hasselhorn et al., 2022, Study 2), we found that the group of participants who had to answer more items (82 items) per measurement occasion had lower within-person variability in momentary mood (but not in state extraversion) and a weaker within-person relationship between state extraversion and momentary mood than the group of participants who had to answer fewer items (33 items) per measurement occasion.[1] Our findings are consistent with the notion that participants in the long questionnaire group responded to items in a more heuristic, less nuanced manner (Fuller-Tyszkiewicz et al., 2013; Podsakoff et al., 2019), and can therefore be interpreted as suggesting that longer questionnaires may lead to compromised data quality in an AA study. In a related vein, previous research on data quality in cross-sectional surveys (Galesic & Bosnjak, 2009) found that questions asked later in a long questionnaire produced lower data quality (as indicated by a lower response rate and lower variability of responses). It can be assumed that longer questionnaires are more cognitively demanding for participants, who may then be motivated to reduce the cognitive demand by responding in a more heuristic way, for example by relying more on RS.

Response styles

RS can be defined as systematic tendencies to prefer specific kinds of response categories over others when answering questionnaire items irrespective of item content (Baumgartner & Steenkamp, 2001; Cronbach, 1946; Paulhus, 1991). As Bolt and Johnson (2009) have argued, RS may reflect participants' attempts to reduce the cognitive demand of distinguishing between levels of agreement, and in line with this idea, Knowles and Condon (1999) found that higher cognitive load increased the magnitude of acquiescence RS.

In the present research, we focused on two of the seven common RS distinguished by Baumgartner and Steenkamp (2001): Midpoint RS (MRS) and Extreme RS (ERS) as response biases. MRS refers to an individual's tendency to prefer the midpoint category of the rating scale, and ERS refers to an individual's tendency to endorse the extreme ends of the rating scale (e.g., Ames & Myers, 2021; Baumgartner & Steenkamp, 2001).

RS can introduce systematic measurement error and thus threaten data quality (Baumgartner & Steenkamp, 2001). Specifically, RS have the potential to explain variability in personality items (Danner et al., 2015), induce differential item functioning (Bolt & Johnson, 2009), distort the factor structure of a multidimensional assessment (Cheung & Rensvold, 2000), bias estimates of the substantive trait intended to be measured (Jin & Wang, 2014), and lead to an overestimation of reliability (Jin & Wang, 2014). Furthermore, RS can confound associations between the substantive trait intended to be measured and other constructs (Bolt & Newton, 2011; Park & Wu, 2019) and threaten construct and predictive validity (Baumgartner & Steenkamp, 2001; van Herk et al.,

[1]Note that the variation in the number of items per measurement occasion was manipulated by using shorter vs. longer versions of the scales (to keep the number of constructs measured equal across experimental groups). However, in order to compare the short vs. long questionnaire groups in terms of within-person variability and the relationship between state extraversion and momentary mood, the analyses were based on those items that were used in both groups (i.e., items that only appeared in the long questionnaire group were excluded from the analyses). Thus, differences in within-person variability and the relationship between the two time-varying variables cannot be attributed to differences in the number of items aggregated into scale scores (and thus not to differences in the reliability of the measures analyzed).

2004). Therefore, it is crucial to account for RS because doing so can increase precision in estimates of the substantive trait and reduce the bias that is associated with RS (Adams et al., 2019; Henninger & Meiser, 2020b), thus maintaining data quality.

Modeling response styles by IRTree approaches

Previous research has proposed a variety of different methods for modeling and accounting for RS, such as count procedures, latent class analytic approaches, or IRT models (Van Vaerenbergh & Thomas, 2013). In recent decades, IRT models have seen an increase in the literature. These models can be divided into threshold-based models and IRTree models (see Böckenholt & Meiser, 2017, for a detailed comparison). Threshold-based models have been extended into multidimensional IRT models (e.g., Bolt & Newton, 2011; Morren et al., 2011), random-threshold models (e.g., Jin & Wang, 2014; Wang & Wu, 2011), or mixture IRT models (e.g., Eid & Rauber, 2000) to account for RS by including additional person parameters or by allowing for population heterogeneity in threshold parameters (see Henninger & Meiser, 2020a, for an overview). IRTree approaches represent an alternative framework for assessing and controlling for RS that combines IRT modeling with decision trees. IRTree approaches model RS as part of a response process (with respect to an ordinal Likert-scale item) by decomposing participants' judgment process into a sequence of binary decisions (Böckenholt, 2012; Böckenholt & Meiser, 2017; De Boeck & Partchev, 2012). Thereby, IRTree models allow researchers to distinguish between processes that are based on the trait of interest and processes that are based on (a priori specified) RS, such as ERS and MRS (Plieninger & Meiser, 2014; Zettler et al., 2016). Böckenholt and Meiser (2017) showed that both groups of models (i.e., threshold-based models and IRTree models) can successfully separate trait-based response processes from RS, and they argued that researchers should choose which method to use to account for RS on the basis of their research question and the requirements of the data in a given situation. As IRTree models require a theory-based decomposition of rating responses into a sequence of decision nodes and the specification of an appropriate statistical model for each node, they are well suited to analyze and control for response style effects in a confirmatory way. Therefore, in the present research, we focused on IRTree models and extended them to multilevel IRTree models to account for the nested data structure of intensive longitudinal data (collected in AA studies).

To decompose the response process into a sequence of decision nodes, IRTree models define a set of dichotomous pseudoitems that are tailored to the Likert-scale format that was used in the study and the RS that were specified a priori (Böckenholt, 2012; Jeon & De Boeck, 2016; Meiser et al., 2019). Figure 1 shows a processing tree diagram for a 5-point Likert item (ranging from 1 to 5), where higher values describe higher agreement with the item content. The processing tree divides the ordinal (Likert-scale) response format into three binary decision nodes. The first decision node refers to the decision of whether a person wants to respond to the midpoint category (which would indicate a neutral response to the item content) of the rating scale or not. If the person chooses the midpoint category (3), the decision process is terminated. Otherwise, the person continues to the next decision node, which reflects the decision to agree or disagree with the item content. In both cases (agreement or disagreement), the person continues to the third decision node, which reflects the decision to respond with an extreme response (1 or 5) or a nonextreme response (2 or 4). In the IRTree model in Figure 1, each of these three decision nodes is captured by a binary pseudoitem, Y_{hvi}, which represents decision node h of person v to item i, where $h = 1, \ldots, 3$; $v = 1, \ldots, N$; and $i = 1, \ldots, I$ (see Table 1).

For each pseudoitem, the probabilities of the possible results can be parametrized in terms of the dichotomous Rasch model, as depicted in the right column of Table 1. The pseudoitems Y_{1vi} are assumed to measure individual differences in MRS (η_1), the pseudoitems Y_{2vi} are assumed to measure individual differences in the substantive trait (θ), and the pseudoitems Y_{3vi} are assumed to measure individual

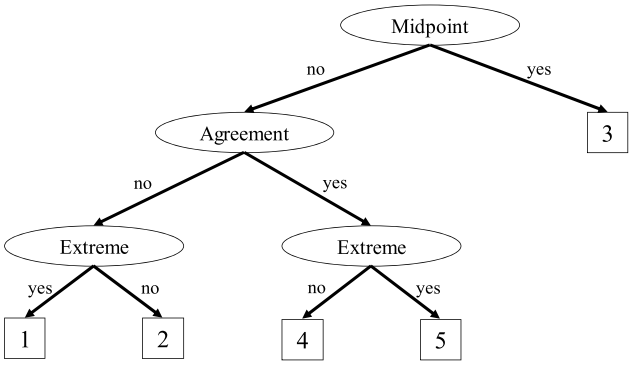

Figure 1. Processing tree diagram of midpoint versus nonmidpoint, agreement versus disagreement, and extreme versus nonextreme responding on a 5-point rating scale (adapted from Böckenholt, 2012).

Table 1. Definition of pseudoitems and node probabilities for the IRTree model in Figure 1.

	Rating category					
	1	2	3	4	5	$p(Y_{hvi} = y_{hvi})$
Midpoint (Y_{1i})	0	0	1	0	0	$\frac{\exp(y_{1vi}(\eta_{1v}-\beta_{1i}))}{1+\exp(\eta_{1v}-\beta_{1i})}$
Agreement (Y_{2i})	0	0	–	1	1	$\frac{\exp(y_{2vi}(\theta_v-\beta_{2i}))}{1+\exp(\theta_v-\beta_{2i})}$
Extreme (Y_{3i})	1	0	–	0	1	$\frac{\exp(y_{3vi}(\eta_{2v}-\beta_{3i}))}{1+\exp(\eta_{2v}-\beta_{3i})}$

differences in ERS (η_2). Note that the agreement and extreme pseudoitems are not defined if a midpoint response is chosen (i.e., if person v selects response category 3 for item i), resulting in missing values for pseudoitems Y_{2i} and Y_{3i}.

A limitation of the traditional type of IRTree model specification is that the substantive trait θ influences only the decision to agree or disagree with the item content and not the decision to respond with an extreme response (1 or 5) or a nonextreme response (2 or 4). To overcome this limitation, Meiser et al. (2019) showed how an ordinal trait-based response process can be integrated into IRTree models by using a multidimensional parametrization of decision nodes. The basic idea of a multidimensional parametrization of decision nodes is that the degree of (dis)agreement, which is assessed using response categories of different intensity (in our example, two categories for disagreement and two categories for agreement), is jointly influenced by a trait-based process and RS. For the tree model for five response categories depicted in Figure 1, assuming an ordinal judgment process implies that the substantive trait θ not only affects whether participants agree or disagree with the item content, but it also influences the fine-grained decision of whether to choose an extreme or a nonextreme response within the disagree and agree categories, respectively. The ordinal judgment process can be implemented by splitting the pseudoitem for extreme responding Y_{3vi} between the categories of disagreement versus agreement (see Table 2) and by specifying the split pseudoitem (i.e., a vs. b) to load on both the trait factor θ and the ERS factor η_2 (see the model equations in the right column of Table 2). The magnitude of the influence of the substantive trait θ on the extreme response decision (i.e., to choose an extreme or a nonextreme response) is represented by the weight α. Thereby, the parameter α allows for a different impact of θ on the overall disagree versus agree decision in pseudoitem Y_{2i} and the more nuanced choice of category. The split pseudoitem Y_{3i} in Table 2 differs only in the direction of the influence of θ: For categories 4 and 5 (agreement

Table 2. Definition of pseudoitems and node probabilities for the two-dimensional parametrization of extreme responding for the IRTree model in Figure 1.

		Rating category					
	Split pseudo item	1	2	3	4	5	$p(Y_{hvi} = y_{hvi})$
Midpoint (Y_{1i})	–	0	0	1	0	0	$\frac{\exp(y_{1vi}(\eta_{1v}-\beta_{1i}))}{1+\exp(\eta_{1v}-\beta_{1i})}$
Agreement (Y_{2i})	–	0	0	–	1	1	$\frac{\exp(y_{2vi}(\theta_v-\beta_{2i}))}{1+\exp(\theta_v-\beta_{2i})}$
Extreme (Y_{3i})	a	1	0	–	–	–	$\frac{\exp(y_{3vi}(\eta_{2v}-\alpha\theta_v-\beta_{3i}))}{1+\exp(\eta_{2v}-\alpha\theta_v-\beta_{3i})}$
	b	–	–	–	0	1	$\frac{\exp(y_{3vi}(\eta_{2v}+\alpha\theta_v-\beta_{3i}))}{1+\exp(\eta_{2v}+\alpha\theta_v-\beta_{3i})}$

categories), participants with a higher (vs. lower) trait value should have a higher probability of selecting the higher (more extreme) response category, whereas for categories 1 and 2, participants with a higher (vs. lower) trait value should have a lower probability of selecting the lower (more extreme) response category. Note that if $\alpha = 0$, the model described in Table 2 is equivalent to the model described in Table 1. For more detailed information about the pseudoitems and node probabilities in IRTree models with unidimensional and two-dimensional node specifications, see Meiser et al. (2019).

Aims of the current research

The overarching aim of the current research was to investigate the impact of (experimentally manipulated) questionnaire length (i.e., the number of items per measurement occasion) in an AA study on RS. In particular, we aimed to estimate individual differences in two substantive traits (state extraversion and state conscientiousness) and participants' preferences for specific kinds of response categories (ERS and MRS) in the two experimental groups (short vs. long questionnaire). To do so while also accounting for the nested data structure (measurement occasions nested in persons), we applied a multigroup multilevel extension of the IRTree model for ordinal and two-dimensional node parametrizations in Table 2. To our knowledge, this is the first study to model RS in an AA study using an IRTree approach (for an application to repeated clinic visits nested in participants using an extension of threshold-based IRT models for ordinal responses, see Deng et al., 2018).

Our preregistered[2] hypothesis was that a longer questionnaire would lead to a greater reliance on RS

[2]The preregistration can be found on the OSF repository (https://doi.org/10.17605/OSF.IO/3V8X6). The hypothesis tested in the present paper is not the only hypothesis that was preregistered in this preregistration document. The reason was that all the hypotheses in this project were preregistered together, but we would have gone beyond the scope of a

(relative to the substantive trait) in an AA study. In terms of the parameters of the IRTree model in Table 2, this means that we expected a longer questionnaire to lead to a smaller influence of the substantive trait θ on the fine-grained decision to choose between an extreme and a nonextreme response, which is reflected by a smaller weight α in the node model of Y_{3i}. Note that a smaller influence of the substantive trait θ (as quantified by the α parameter; see Table 2) corresponds to a larger relative impact of ERS on the choice of (non)extreme agreement and disagreement categories, respectively.

As an application of the multilevel IRTree model, and to test our hypothesis about the effect of questionnaire length in an AA study on the relative impact of RS, we used a dataset (Hasselhorn et al., 2022, Study 2) that has previously been used to examine the effect of experimentally manipulated questionnaire length on data quantity (i.e., compliance), perceived burden, and aspects of data quality (intraindividual variability, within-person relationship between time-varying variables). This dataset has not yet been used to model RS.

In addition to testing our hypothesis about the effect of questionnaire length on the relative impact of RS, we conducted a series of supplemental exploratory analyses to explore the effects that RS had in our AA data (and the differences in these effects between experimental groups). In these exploratory analyses, we compared models that accounted for RS in the AA data with models that did not account for RS in the AA data (for details, see Data Analytic Models section).

Method

Study design

The study consisted of an initial online survey (assessing demographic variables and trait self-report measures), an AA phase across 14 days with measurement occasions per day (with a short or long questionnaire, depending on the experimental condition that participants had been randomly assigned to), and a retrospective online survey (assessing trait self-reports again, as well as retrospective measures that were unrelated to the present research).

In the AA phase, the short questionnaire group had to answer 33 items (or 36 items in the evening) per questionnaire, and the long questionnaire group had to answer 82 items (or 85 items in the evening). The average response time for one questionnaire in the short questionnaire group ($M = 1.64$ min, $SD = 0.63$) was lower, on average, than in the long questionnaire group ($M = 3.91$ min, $SD = 3.43$). The two groups answered questions about the same substantive constructs (at each occasion: momentary motivation, time pressure, state personality, situation characteristics, and momentary mood; additionally, at the last occasion of the day: perceived burden due to study participation). This allowed us to investigate the effect of questionnaire length without the confounding effect of measuring different substantive constructs between the groups. The difference in the number of items between these groups was achieved by using a short versus a long version for most of the measures of the constructs. The constructs that were measured with fewer items in the short questionnaire group compared with the long questionnaire group were situation characteristics (8 vs. 32 items), pleasant-unpleasant mood (2 vs. 4 items), calm-tense mood (1 vs. 2 items), wakefulness-tiredness (1 vs. 2 items), and state openness to experience, agreeableness, and neuroticism (1 vs. 8 items). The state extraversion and state conscientiousness constructs, which were used to model RS in the present research, were measured with the same number of items (8 items per construct) across experimental groups.

Participants

Participants were required to be currently enrolled as a student, to be in possession of a smartphone, to speak German, and to be at least 18 years old. Participants were recruited *via* flyers, e-mails, and posts on Facebook in January 2020, and the last questionnaire was sent to participants on February 10, 2020.

A total of 303 individuals filled out the initial online survey, 284 individuals took part in the AA phase that followed (143 individuals in the short questionnaire condition), and 235 individuals responded to the retrospective online survey after the AA phase (within the prespecified time frame of 5 days). Participants who did not respond to the retrospective online survey were not excluded from the analyses.

single paper if we had attempted to test/report all the hypotheses at once. Some of the other preregistered hypotheses have been (or will be) tested/reported in separate papers (e.g., Hasselhorn et al., 2022). The data analytic models used to test the current hypothesis deviate from the preregistered data analytic models. At the time of the preregistration (January, 2020), we had planned to aggregate the data across days to analyze RS in AA data. Only after the preregistration did we find out that it might be possible to apply IRTree models to AA data using an MSEM framework. Additionally, we chose not to use a partial credit tree model because doing so would have confounded the substantive trait we intended to measure and ERS.

The final sample consisted of 284 students (short questionnaire group: 83% women; age: $M = 23.19$, $SD = 3.44$, Range = 18 to 39 years, 4457 completed measurement occasions; long questionnaire group: 87% women; age: $M = 22.91$, $SD = 3.80$, Range = 18 to 55 years, 4214 completed measurement occasions).[3]

Procedure

All study procedures were approved by the psychological ethics committee at the University Koblenz-Landau (ethics approval number 228). After obtaining informed consent, the study began with an initial online survey to assess trait measures and sociodemographic information. Subsequently, participants were randomly assigned to one of two experimental conditions (short questionnaire or long questionnaire) and were informed about the upcoming AA phase at least 2 days in advance. The AA phase of 14 days began on the next possible Monday or Thursday. All participants received three links to questionnaires *via* SMS per day (10:00, 14:00, and 18:00) and had 45 min until they could no longer start the questionnaire. After the 14-day AA phase, participants received a link to the retrospective online survey *via* SMS. This online survey had to be completed within a 5-day time frame. Participants received up to 30€in exchange for their participation depending on their compliance rate (25% = 3€, 50% = 10€, 75% = 20€, and 90% = 30€). Furthermore, when they filled out the initial online survey, they could choose to receive personal feedback on the measured constructs after they participated. In the short questionnaire group (long questionnaire group) 134 (131) participants requested feedback, and 9 (10) participants did not want feedback.

Measures

Questionnaire length

We included a dummy-coded questionnaire length factor, with a value of 0 for the short questionnaire and 1 for the long questionnaire.

State extraversion and conscientiousness

We measured state extraversion and state conscientiousness with an adapted version of the adjectives from Saucier's (1994) unipolar Big Five Mini-Markers (Comensoli & MacCann, 2015). Participants indicated how they had *behaved in the last half hour* on eight items for state extraversion (bashful [reverse-scored], bold, energetic, extraverted, quiet [reverse-scored], shy [reverse-scored], talkative, and withdrawn [reverse-scored]) and on eight items for state conscientiousness (careless [reverse-scored], disorganized [reverse-scored], efficient, inefficient [reverse-scored], organized, practical, sloppy [reverse-scored], systematic, creative, unenvious, unsympathetic). The response format was a 5-point Likert scale with each pole labeled (1 = *extremely inaccurate* to 5 = *extremely accurate*). A higher score indicated more extraverted (or more conscientious) behavior. For state extraversion, the within-person ω (Geldhof et al., 2014) was 0.72, and the between-person ω was .59. For state conscientiousness, the within-person ω was 0.79, and the between-person ω was .80.[4]

Momentary pleasant-unpleasant mood

We measured momentary pleasant-unpleasant mood with an adapted short version of the Multidimensional Mood Questionnaire (Steyer et al., 1997), which has been used in previous AA studies (Lischetzke et al., 2012; Ottenstein & Lischetzke, 2020). We used two items from the adapted short version in which participants indicated how they *felt at the moment* on two items (bad-good [reverse-scored], unwell-well). The response format was a 7-point Likert scale with each pole labeled (e.g., 1 = *very unwell* to 7 = *very well*). A higher score indicated more pleasant mood. The within-person ω (Geldhof et al., 2014) was 0.84, and the between-person ω was 0.97.

Global self-report of personality measured in the initial online survey

We measured global self-report of extraversion and conscientiousness with unipolar adjective scales (Trierweiler et al., 2002) that had four adjectives per dimension in the initial online survey. Participants indicated how they *best identified as a person* on each adjective. The response format was a 5-point Likert

[3]As an additional randomization check, we tested whether the two experimental groups differed on age, gender, pleasant-unpleasant mood, and global self-report of personality (Big 5) as measured in the initial online survey. We corrected for multiple testing using Benjamini and Hochberg (1995) procedure for controlling the false discovery rate (FDR) in these eight tests. None of the variables were significantly different between groups after the FDR was corrected.

[4]The between-person ω for state extraversion appears to be relatively low (e.g., compared to the between-person ω for conscientiousness). Analyzing the item correlations, we identified items (e.g., bashful [reverse-scored], or shy [reverse-scored]) with relatively low item correlations (both between- and within persons) within the extraversion scale. However, these items are an important part of the theory-driven construct of extraversion, as they may measure distinct facets of the construct that were not frequently observed in our student sample. Therefore, we believe that it is important not to exclude these items, as this would change the interpretation of extraversion as a construct.

scale ranging from 1 (*not at all*) to 5 (*very much so*). We calculated a mean score across the four items (extraversion: sociable, companionable, vivacious, and spirited; conscientiousness: industrious, diligent, dutiful, and ambitious) in each dimension such that a higher value indicated a higher standing on the respective personality trait. Revelle's omega total (McNeish, 2018) was 0.73 for global self-report of extraversion and 0.82 for global self-report of conscientiousness.

Data analytic models

Selection of an MSEM IRTree base model

We conducted a series of multigroup IRTree models in a Multilevel Structural Equation Modeling (MSEM) framework on the data from the AA phase of the study to test the effect of experimentally manipulated questionnaire length on relative RS effects. Specifically, we used the processing tree model described in Figure 1 and the parametrization of pseudoitems and node probabilities described in Tables 1 and 2 to convert each of the eight Likert-scale items for each construct (state extraversion and state conscientiousness) into a sequence of three pseudoitems (for a total of 24 pseudoitems for each construct). Table 1 describes an IRTree model with unidimensional node parameterizations, and Table 2 describes an IRTree model with two-dimensional node specifications for (non-)extreme responding. The two-dimensional parametrization of the pseudoitems Y_{3i} resembles a bifactor model (Eid et al., 2017) in which the midpoint pseudoitems Y_{1i} and the agreement pseudoitems Y_{2i} each load on one factor (θ for the agreement pseudoitems and η_1 for the midpoint pseudoitems), whereas the extreme responding pseudoitem Y_{3i} loads on both θ and η_2. That is, after converting the Likert-scale items into pseudoitems with the two-dimensional parametrization, we used eight midpoint pseudoitems Y_{1i}, eight agreement pseudoitems Y_{2i}, and 16 split extreme pseudoitems Y_{3i} for each construct in the bifactor model structure (for a total of 32 pseudoitems and split pseudoitems for each substantive construct). The described bifactor model structure for both constructs can be seen in the upper part of the model in Figure 2, where items $i = 1, \ldots, 8$ measured state extraversion, and items $i = 9, \ldots, 16$ measured state conscientiousness.

To account for the experimental design and the multilevel data structure (measurement occasions nested within persons), we extended the IRTree models to multigroup multilevel IRTree models and specified them in a multigroup multilevel structural equation modeling (MSEM) framework in MPlus (Muthén & Muthén, 1998-2022). The MSEM model for the two-dimensional parametrization of the extreme responding pseudoitems at the between- and within-person levels can be seen in Figure 2. Note that the covariances at the between-person and within-person levels are not included in the figure. To account for the multilevel data structure, the subscript $t = 1, \ldots, T$ represents measurement occasions (that are nested within persons) so that the response process for the original Likert scale item Y_{tvi} of person v in measurement occasion t to item i became conceptualized as a set of pseudoitems Y_{htvi}. Within the multigroup MSEM framework, we modeled each substantive construct (extraversion and conscientiousness), MRS (η_1), and ERS (η_2) at the measurement occasion (within-person) level and at the person (between-person) level. In the following, to distinguish between the latent constructs at the different levels, we refer to the latent constructs at the between-person level as *traits* and to the latent constructs at the within-person level as *states*, a practice that is in line with theoretical accounts of within- and between-person differences in personality dimensions (e.g., Fleeson, 2001; Fleeson & Jayawickreme, 2015).

To test whether the experimental groups differed in the α parameter for a substantive trait at the between-person level, which quantifies the influence of the substantive trait θ on the fine-grained decision between an extreme and a nonextreme response, we first determined the optimal model that fit the data best in each experimental group separately, before analyzing differences in the α parameters across the two experimental groups. In step 1, within each experimental group, we fixed α to zero (which is equivalent to the unidimensional node parameterizations) at the within-person level and determined whether a uni- versus a two-dimensional structure (see the model equations in Tables 1 and 2) held at the between-person level and whether - for a two-dimensional parametrization – the α parameters could be set equal across the two substantive constructs (i.e., extraversion and conscientiousness; Models 1 to 3 in Table 3). In the second step, we fixed the person-level dimensional structure (uni- vs. two dimensional parametrization) according to the results from step 1 and additionally scrutinized whether a more complex (two-dimensional) structure was needed at the within-person level (Models 4 to 7 in Table 3). For each model (Models 1 to 7), we estimated means for each latent variable at both levels (θext_v, θconc_v, η_{1v}, η_{2v}, θext_{vj}, θconc_{vj}, η_{1vj}, η_{2vj}, see Figure 2). We used the Akaike information criterion

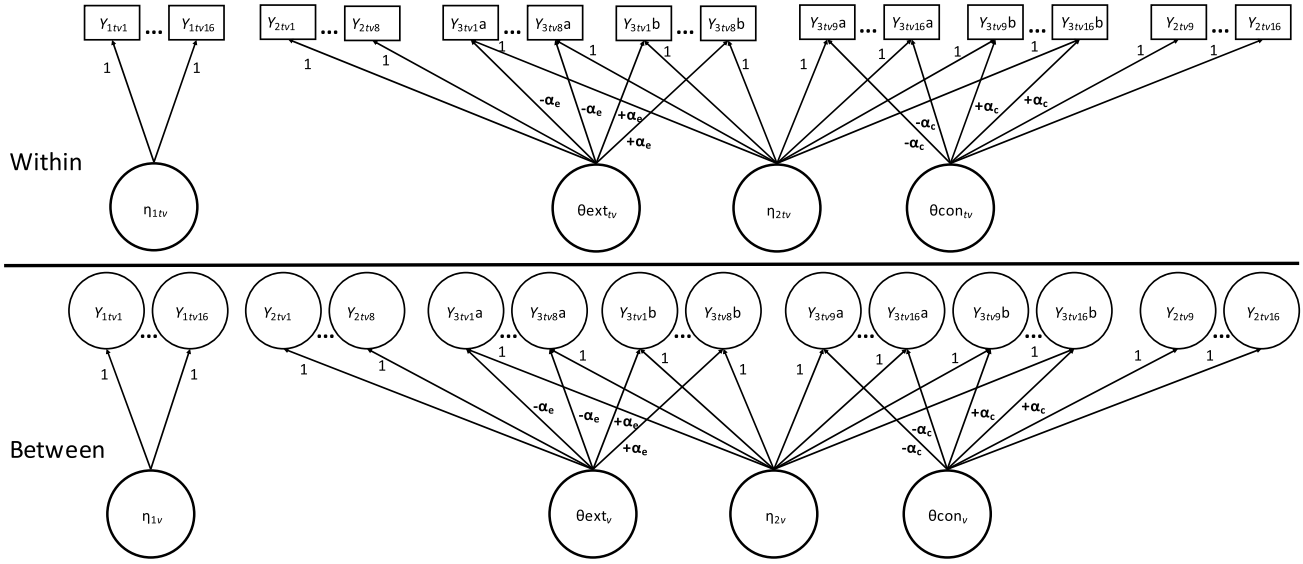

Figure 2. Multilevel structural equation model for the two-dimensional parametrization of extreme responding at the between-person level and the within-person level for one experimental group.
Note. θext = factor for extraversion; θcon = factor for conscientiousness; η_1 = factor for midpoint response style; η_2 = factor for extreme response style; Y_{htvi} represents decision node h of person v within measurement occasion t to item i. For $i = 1, \ldots, 8$, the items measured state extraversion, and for $i = 9, \ldots, 16$, the items measured state conscientiousness. α_e represents the α parameter for extraversion and α_c represents the α_e parameter for conscientiousness. The α parameters α_e and α_c are estimated separately on each level. Covariances between latent variables at the between-person level and the within-person level are not displayed.

Table 3. Model fit statistics for IRTree models in a multilevel structural equation modeling framework with different α parameter at the between-person level and the within-person level.

	Within-person level			Between-person level					
	unidim	two-dim		unidim	two-dim				
Model	$\alpha = 0$	equal α	freely est. α	$\alpha = 0$	equal α	freely est. α	AIC	BIC	Npar
Short questionnaire group									
Model 1	x			x			185201.592	185662.553	72
Model 2	x				x		184731.486	185198.849	73
Model 3	**x**					**x**	**184667.505**	**185141.270**	**74**
Model 4		x				x	184952.151	185432.318	75
Model 5			x			x	184975.008	185461.577	76
Long questionnaire group									
Model 1	x			x			168572.058	169028.982	72
Model 2	x				x		168307.029	168770.299	73
Model 3	x					x	168308.819	168778.435	74
Model 6		x			x		168306.907	168776.523	74
Model 7			**x**		**x**		**168256.118**	**168732.081**	**75**

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; Npar = number of parameters; unidim = unidimensional parametrization of the pseudoitems; two-dim = two-dimensional parametrization of the pseudoitems; equal $\alpha = \alpha$ parameters set equal across the two substantive constructs (extraversion, conscientiousness); freely est. $\alpha = \alpha$ parameters freely estimated across the two substantive constructs (extraversion, conscientiousness); x = selected parametrization of the pseudoitems on the respective level; Bold indicates the best model for each group.

(AIC) and the Bayesian information criterion (BIC; Burnham & Anderson, 2004) as model selection criteria, with lower values indicating a better balance of model fit and parsimony to assure suitability of the model to the empirical data while avoiding overparameterizing complex models.[5]

[5]We acknowledge that we did not specify a priori how we would proceed if the AIC and the BIC led to different conclusions. However, this has not been the case.

Effects of questionnaire length

After determining the model that fit the data best in each experimental group separately, we specified multigroup MSEM IRTree models. At the between-person and within-person levels, we used parametrizations that were as parsimonious as possible and as complex as needed (as indicated by the results of the single-group models). To ensure that potential differences in the α parameters across groups were not caused by differences in the underlying structural model, the

same model structure was specified for both experimental groups. To test whether the two experimental groups differed in the α parameters across experimental groups, we compared two multigroup MSEM IRTree models: In one model, the α parameters were constrained to be equal across experimental groups (at the between-person level), and in the other model, the α parameters were freely estimated for each experimental group (at the between-person level). Subsequently, we compared the model fit of the constrained and unconstrained (free) models using the AIC and the BIC. Note that the chosen (parsimonious as possible and as complex as needed) model determined how the α parameters were estimated in the constrained and the unconstrained model (whether the α parameters were freely estimated across the two substantive constructs at both levels).[6]

Supplemental exploratory analyses

To explore the effects that RS had in our AA data and the differences in these effects between experimental groups, we conducted a series of supplemental exploratory analyses. To estimate the effects that RS had in our AA data, we compared models that accounted for RS in the AA data (as the models described above) with models that did not account for RS in the AA data. To quantify the differences between the two types of models (models with RS and models without RS), we computed regression analyses between the previously described substantive constructs (state extraversion and state conscientiousness) and two (sets of) external criteria: momentary pleasant-unpleasant mood, which was measured in the AA phase of the study, and global self-report of extraversion and conscientiousness, which were measured in the initial online survey. Specifically, we estimated the regression of momentary pleasant-unpleasant mood on state extraversion at both the within-person and between-person levels and the regression of extraversion and conscientiousness (measured in the AA phase) on global self-report of extraversion and conscientiousness (measured in the initial online survey) at the between-person level. We corrected for multiple testing using the procedure presented by Benjamini and Hochberg (1995) for each set of external criteria. To explore whether the experimental groups differed in these regression coefficients, we freely estimated the regression coefficients for each experimental group in a multigroup model. We used the model that was the best fitting model in the multigroup MSEM IRTree models (described above) as the multigroup model that accounted for RS and subsequently added the external criteria. Details on the supplemental exploratory analyses (e.g., restrictions) can be found in the Supplemental Online Material (https://doi.org/10.17605/osf.io/xt3rf). All models were computed in Mplus (Muthén & Muthén, 1998-2022).

[6]The α parameters were not constrained across the (measurement occasion and person) levels across all models (Models 1 to 7). We estimated equal variances and covariances at the within-person and the between-person level across experimental groups. Within each experimental group, the means for each substantive trait (θext_v, $\theta conc_v$) were estimated freely so that differences in the α parameters could be directly compared across groups.

Results

All MSEM IRTree models were applied to the observed 8,671 measurement occasion, which were nested in 284 participants.

MSEM IRTree models for the short questionnaire group

To identify the best fitting model in the short questionnaire group, we determined which dimensional structure (uni- vs. two-dimensional) held at the between-person level and if the α parameters could be set equal across the two substantive constructs (i.e., extraversion and conscientiousness; Models 1 to 3 in Table 3). The upper panel of Table 3 shows the model fit statistics for the short questionnaire group. According to the BIC and AIC, the best fitting model was Model 3, which used the two-dimensional parametrization of the pseudoitems at the between-person level and freely estimated α parameters for the two substantive constructs.

In the second step, we determined whether a more complex (two-dimensional) structure was needed at the within-person level (Models 4 and 5 in Table 3), whereas we fixed the dimensional structure at the between-person level according to the result of the first step (Model 3). The upper panel of Table 3 shows the model fit statistics. According to the BIC and AIC, the best fitting model was Model 3, which used the unidimensional parametrization at the within-person level. Therefore, Model 3 was the best fitting model for the short questionnaire group, which used the two-dimensional parametrization of the pseudoitem for extreme responding at the between-person level with freely estimated α parameters for the two substantive constructs and the unidimensional parametrization at the within-person level.

MSEM IRTree models for the long questionnaire group

To identify the best fitting model in the long questionnaire group, we determined which dimensional structure (uni- vs. two-dimensional) held at the between-person level and if the α parameters could be set equal across the two substantive constructs (i.e., extraversion and conscientiousness; Models 1 to 3 in Table 3). The lower panel of Table 3 shows the model fit statistics for the long questionnaire group. According to the BIC and AIC, the best fitting model was Model 2, which used the two-dimensional parametrization of the pseudoitems at the between-person level with α parameters set equal across the two substantive constructs.

In the second step, we determined whether a more complex (two-dimensional) structure was needed at the within-person level (Models 6 and 7 in Table 3), whereas we fixed the dimensional structure at the between-person level according to the result of the first step (Model 2). The lower panel of Table 3 shows the model fit statistics. According to the BIC and AIC, the best fitting model was Model 7, which used the two-dimensional parametrization of the pseudoitems at the within-person level with freely estimated α parameters for the two substantive constructs. Therefore, Model 7 was the best fitting model for the long questionnaire group, which used the two-dimensional parametrization of the pseudoitems at both levels with α parameters set equal across the two substantive constructs at the between-person level (traits) and freely estimated α parameters for the substantive constructs at the within-person level (states).

Differences in RS across experimental groups

As Model 3 was the best fitting model in the short questionnaire group and Model 7 was the best fitting model in the long questionnaire group, we selected the two-dimensional parametrization at both the within-person level and the between-person levels, with freely estimated α parameters for the two substantive constructs at both levels in the multigroup MSEM IRTree models. Figure 2 displays the final structural model. Note that for readability, only one experimental group is depicted (the same parametrization was used in the other experimental group). Covariances at the between-person and within-person levels are not included in the figure. To analyze differences in RS across the two experimental groups (which are quantified by the α parameters), we compared the unconstrained model (with freely estimated α parameters at the between-person level across experimental groups) with a constrained model that had α parameters (for each substantive trait at the between-person level) that were set to be equal across experimental groups. The unconstrained model (AIC = 354012.155, BIC = 354598.777) fit the data better than the constrained model (AIC = 354077.139, BIC = 354635.490). This result means that the two experimental groups differed with regard to the influence of the substantive trait θ_v on the fine-grained decision between an extreme and a nonextreme response. To investigate the direction of the effect of questionnaire length, we compared the α parameters for each substantive trait. In line with our hypothesis, the α parameter for trait extraversion was smaller in the long questionnaire group ($\alpha = 0.277$, $SE = 0.02$) than in the short questionnaire group ($\alpha = 0.402$, $SE = 0.02$). Similarly, the α parameter for trait conscientiousness was smaller in the long questionnaire group ($\alpha = 0.189$, $SE = 0.01$) than in the short questionnaire group ($\alpha = 0.438$, $SE = 0.02$). These findings show that the relative impact of the trait was smaller, and that of ERS stronger, in the condition with longer questionnaires per measurement point.

Supplemental exploratory analyses: relationships between substantive constructs in models adjusting for RS and in models not adjusting for RS

We additionally explored whether the estimated relationship between substantive constructs was similar across models that accounted for RS (using the IRTree approach) and models that did not account for RS ("standard" MSEM models). Results of these exploratory analyses can be found in the Supplemental Material. Here, we briefly summarize the pattern of results that emerged.

Relationship between state extraversion and momentary mood

First, we analyzed the relation between state extraversion (measured in the AA) and momentary pleasant-unpleasant mood (also measured in the AA) on the within- and the between-person level. In both experimental groups, state extraversion was positively related to pleasant-unpleasant mood at the within-person level and at the between-person level when RS were accounted for. When RS were *not* accounted for, the same pattern of results emerged (i.e., a positive significant relationship on both levels). With respect to the size of the regression coefficients, three out of

four coefficients were descriptively greater in the models that accounted for RS compared with the models that did not account for RS.

Relationships of extraversion and conscientiousness as measured in AA with global self-reports

Second, we analyzed the relation between extraversion (measured in the AA phase) and the global self-report of extraversion (measured in the initial online survey), and the relation between conscientiousness (measured in the AA phase) and the global self-report of conscientiousness (measured in the initial online survey) at the between-person level. In both experimental groups, when RS were accounted for, and when RS were *not* accounted for, individual differences in extraversion were significantly positively related to the global self-report of extraversion, and individual differences in conscientiousness were significantly positively related to the global self-report of conscientiousness. With respect to the size of the regression coefficients, three out of four coefficients were descriptively smaller in the models that accounted for RS compared with the models that did *not* account for RS.

Discussion

The aim of the current study was to investigate the impact of questionnaire length on the relative effects of traits and RS in an AA study, as RS are a potential threat to the data quality of AA studies. To test whether a longer questionnaire would lead to a greater effect of RS (relative to the substantive trait) in an AA study, we used multigroup multidimensional IRTree models in an MSEM framework. In line with our expectations, our main finding was that, in the group with the long (vs. the short) questionnaire, there was less of an influence of the substantive trait on the fine-grained decision between an extreme and a nonextreme response. That is, as expected, the responses of participants in the long (vs. the short) questionnaire group were influenced more strongly by RS relatively to the trait.

Our finding of an increased (relative) impact of RS on responses when participants are required to answer more items per measurement occasion is consistent with previous results showing negative effects of increased questionnaire length on indicators of data quality in AA (Eisele et al., 2022) and in cross-sectional surveys (Galesic & Bosnjak, 2009). Previous analyses of the present dataset (Hasselhorn et al., 2022, Study 2) found that longer questionnaires were associated with a smaller degree of intraindividual variability in one of two constructs and a weaker within-person relationship between the two constructs. The analyses and results presented in the present paper add to this picture by showing (more directly) that questionnaire length had an impact on the response process (i.e., on the way participants select among response categories when answering AA items in their daily life).

It is important to note that the experimental manipulation of questionnaire length in the study by Eisele et al. (2022) and in our study (Hasselhorn et al., 2022, Study 2, and the present research) included a similar number of items for the short questionnaire group (30 items in the study by Eisele et al., 2022, and 33 items in our study). For the long questionnaire group, participants in our study answered more items per occasion (82 items) than participants in the study by Eisele et al. (60 items). On the basis of these two studies, it is not possible to identify the threshold (in terms of number of items) at which the changes in (aspects of) data quality occur and it remains an open question whether these changes might be influenced by factors other than questionnaire length, such as the complexity of the items, item length (e.g., the number of words in each item), the cognitive load involved in answering each item, and the software used to measure the items. Future research should investigate the optimal number of items in an AA study (i.e., the number of items that participants can manage without aspects of data quality becoming impaired) and investigate the potential interactions between other factors that might influence the optimal number of items.

With respect to the psychological process(es) behind the differential effects of questionnaire length on data quality in an AA study, we assumed that a longer (vs. a shorter) questionnaire leads to higher cognitive load for participants. This assumption was based on the arguments made by Bolt and Johnson (2009) and Knowles and Condon (1999), who argued that higher cognitive load would lead to a larger magnitude of acquiescence) RS (Knowles and Condon (1999) and that RS may reflect participants' attempts to reduce cognitive demand (Bolt & Johnson, 2009). It seems obvious that questionnaires with more items (vs. fewer items) would place participants under greater cognitive load as they attempted to complete such a questionnaire. Specifically, they have to complete each component of the response process, comprehension of the item, retrieval of relevant information, use of that information to make

necessary judgments, and selection and reporting of a response (see Tourangeau et al., 2000) more times per questionnaire with a long questionnaire (vs. a short questionnaire). However, we did not directly measure cognitive load in our study; thus, we cannot rule out alternative explanations of underlying psychological processes. Therefore, future research is needed to demonstrate whether the effect of questionnaire length on the magnitude of RS is driven by the cognitive load that the questionnaire imposes on participants as they complete the questionnaire.

To our knowledge, the current study is the first to model RS in an AA study using IRTree models. IRTree models are used to separate latent judgment processes that are based on the substantive trait from effects of RS. Furthermore, our chosen multigroup MSEM IRTree model allowed us to account for the nested data structure in AA studies (measurement occasions nested in persons), model the substantive states, substantive traits, and RS simultaneously, and investigate the effect of questionnaire length (as a between-person experimental factor) on RS. Our results are in line with research by Meiser et al. (2019), who found that the multidimensional parametrization of the node probabilities better described the latent judgment process compared with the unidimensional parametrization. Specifically, we found that the substantive trait influences not only participants' decision about whether they generally agree or disagree with the item content but also the fine-grained decision to choose between an extreme and a nonextreme response category of agreement or disagreement. We recommend that researchers who want to investigate RS using IRTree models in an AA study use our modeling approach to account for the nested data structure within AA and to (better) capture the latent judgment processes.

With respect to the supplemental exploratory analyses in which we descriptively compared models that accounted for RS with models that did not account for RS, the size of regression coefficients varied (seemingly) unsystematically across these two types of models. In some analyses, the relationships between variables were descriptively greater in the models that accounted for RS compared with the models that did not account for RS, and in other analyses, the relationships between analyzed variables were descriptively smaller in the models that accounted for RS compared with the models that did not account for RS. The latter pattern (i.e., inflated correlations between different constructs if response styles are not controlled for) has been described in previous research on RS (Böckenholt & Meiser, 2017). Due to the exploratory nature of the analyses, we refrain from speculating about potential reasons for the mixed results. More research is needed to elucidate the effects of RS on AA data because there might be other conditions under which RS bias the substantive interpretation between variables (e.g., different substantive constructs, design characteristics, or sample characteristics).

Limitations

Limitations need to be considered when interpreting the current findings. To investigate the current hypothesis, we designed our study in such a way that we could acquire a relatively large data set to ease the convergence of our (relatively) complex model. Specifically, we used eight items to measure each substantive construct (state extraversion and state conscientiousness) three times a day for 14 days in the AA phase. Note that we chose substantive constructs that typically show relatively weak intercorrelations to further ease the convergence of the model. Other AA studies usually measure substantive constructs with fewer items per construct (with some studies using only one or two items per construct). Additionally, other AA studies might use fewer questionnaires per day or a shorter AA phase, resulting in a smaller total number of questionnaires per participant. These factors decrease the information available in the data set and might lead to (convergence) problems when estimating (multigroup) MSEM IRTree models. However, we do not know the boundary conditions that have to be met so that the proposed MSEM IRTree can be estimated. Future research should investigate the boundary conditions that have to be met so that model estimation can be ensured for nested data sets.

In our study, we manipulated one of many central design choices in an AA study. However, many other design choices (e.g., the sampling frequency or the number of days used to survey people) that might affect RS or other aspects of the data quality have yet to be explored. Additionally, there might be interactions between design choices that influence the impact of questionnaire length on RS or other (aspects of) data quality. For instance, the effect of questionnaire length on RS might diminish when a smaller number of days is used in the AA phase (e.g., 3 days instead of the 2 wk we used). Future research should investigate other design choices in an AA study and possible interactions between design choices on their effects on RS (and other aspects of data quality).

Another limitation is the composition of our (student) sample (participants who were young and highly

educated, with a large proportion of women), which might restrict the generalizability of our findings. We do not know whether the findings of the current study depended on certain characteristics of our sample. For example, it is possible that women have a higher baseline ERS than men (Batchelor & Miao, 2016), or that the effect of questionnaire length on RS depends on age (with a more pronounced effect in older participants), or intelligence (with a less pronounced effect in more intelligent participants). Future research might analyze whether the effects of longer AA questionnaires differ across different samples.

Conclusions

The present research is the first to analyze the impact of questionnaire length on RS in an AA study. By extending IRTree models to a (multigroup) MSEM framework, we also presented a promising modeling approach that can be applied to account for RS in nested data such as AA data. We found that a longer (vs. a shorter) questionnaire, operationalized as the number of items per measurement occasion, led to a greater magnitude of RS in our AA study. Based on the results of the current research, we suggest that researchers should avoid a large number of items per measurement occasion in order to reduce the effects of RS relative to the effects of the substantive traits that are the focus of measurement. Although further validation of our findings is necessary, we hope that researchers will consider our findings when planning an AA study in the future.

Author information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant GRK 2277 (Research Training Group "Statistical Modeling in Psychology") from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *The British Journal of Mathematical and Statistical Psychology*, *72*(3), 466–485. https://doi.org/10.1111/bmsp.12169

Ames, A. J., & Myers, A. J. (2021). Explaining variability in response style traits: A covariate-adjusted IRTree. *Educational and Psychological Measurement*, *81*(4), 756–780. https://doi.org/10.1177/0013164420969780

Arslan, R. C., Reitz, A. K., Driebe, J. C., Gerlach, T. M., & Penke, L. (2020). Routinely randomize potential sources of measurement reactivity to estimate and adjust for biases in subjective reports. *Psychological Methods*, *26*(2), 175–185. https://doi.org/10.1037/met0000294

Batchelor, J., & Miao, C. (2016). Extreme response style: A meta-analysis. *Journal of Organizational Psychology*, *16*(2), 51–62.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U., & 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

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Bolt, D. M., & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

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

Burnham, K. P., & Anderson, D. R. (Eds.) (2004). *Model selection and multimodel inference*. Springer. https://doi.org/10.1007/b97636

Carpenter, R. W., Wycoff, A. M., & Trull, T. J. (2016). Ambulatory assessment: New adventures in characterizing dynamic processes. *Assessment*, *23*(4), 414–424. https://doi.org/10.1177/1073191116632341

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212. https://doi.org/10.1177/0022022100031002003

Comensoli, A., & MacCann, C. (2015). Emotion appraisals predict neuroticism and extraversion: A multilevel investigation of the appraisals in personality (AIP) model. *Journal of Individual Differences*, *36*(1), 1–10. https://doi.org/10.1027/1614-0001/a000149

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

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

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

Deng, S., McCarthy, D. E., Piper, M. E., Baker, T. B., & Bolt, D. M. (2018). Extreme response style and the measurement of intra-individual variability in affect. *Multivariate Behavioral Research*, *53*(2), 199–218. https://doi.org/10.1080/00273171.2017.1413636

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20–30. https://doi.org/10.1027//1015-5759.16.1.20

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Fahrenberg, J. (2006). Assessment in daily life. A review of computer-assisted methodologies and applications in psychology and psychophysiology, years 2000—2005. Retrieved from http://www.ambulatory-assessment.org/

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: Traits as density distributions of states. *Journal of Personality and Social Psychology*, *80*(6), 1011–1027. https://doi.org/10.1037/0022-3514.80.6.1011

Fleeson, W., & Jayawickreme, E. (2015). Whole trait theory. *Journal of Research in Personality*, *56*, 82–92. https://doi.org/10.1016/j.jrp.2014.10.009

Fuller-Tyszkiewicz, M., Skouteris, H., Richardson, B., Blore, J., Holmes, M., & Mills, J. (2013). Does the burden of the experience sampling method undermine data quality in state body image research? *Body Image*, *10*(4), 607–613. https://doi.org/10.1016/j.bodyim.2013.06.003

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2022). The effects of assessment intensity on participant burden, compliance, within-person variance, and within-person relationships in ambulatory assessment. *Behavior Research Methods*, *54*(4), 1541–1558. https://doi.org/10.3758/s13428-021-01683-6

Hasselhorn, K., Ottenstein, C., & Lischetzke, T. (2023). Modeling careless responding in ambulatory assessment studies using multilevel latent class analysis: Factors influencing careless responding. *Psychological Methods*, Advance online publication. https://doi.org/10.1037/met0000580

Henninger, M., & Meiser, T. (2020a). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, *25*(5), 560–576. https://doi.org/10.1037/met0000249

Henninger, M., & Meiser, T. (2020b). Different approaches to modeling response styles in divide-by-total item response theory models (part 2): Applications and novel extensions. *Psychological Methods*, *25*(5), 577–595. https://doi.org/10.1037/met0000268

Himmelstein, P. H., Woods, W. C., & Wright, A. G. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychological Assessment*, *31*(7), 952–960. https://doi.org/10.1037/pas0000718

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*(3), 1070–1085. https://doi.org/10.3758/s13428-015-0631-y

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*(1), 116–138. https://doi.org/10.1177/0013164413498876

Knowles, E. S., & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*(2), 379–386. https://doi.org/10.1037/0022-3514.77.2.379

Lischetzke, T., Pfeifer, H., Crayen, C., & Eid, M. (2012). Motivation to regulate mood as a mediator between state extraversion and pleasant–unpleasant mood. *Journal of Research in Personality*, *46*(4), 414–422. https://doi.org/10.1016/j.jrp.2012.04.002

May, M., Junghaenel, D. U., Ono, M., Stone, A. A., & Schneider, S. (2018). Ecological Momentary Assessment Methodology in Chronic Pain Research: A Systematic Review. *The Journal of Pain*, *19*(7), 699–716. https://doi.org/10.1016/j.jpain.2018.01.006

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412–433. https://doi.org/10.1037/met0000144

Mehl, M. R., & Conner, T. S. (Eds.) (2012). *Handbook of research methods for studying daily life*. Guilford.

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *The British Journal of Mathematical and Statistical Psychology*, *72*(3), 501–516. https://doi.org/10.1111/bmsp.12158

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with extreme response style in cross-cultural research: A restricted latent class factor analysis approach. *Sociological Methodology*, *41*(1), 13–47. https://doi.org/10.1111/j.1467-9531.2011.01238.x

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Ottenstein, C., & Lischetzke, T. (2020). Development of a novel method of emotion differentiation that uses open-ended descriptions of momentary affective states. *Assessment*, *27*(8), 1928–1945. https://doi.org/10.1177/1073191119839138

Park, M., & Wu, A. D. (2019). Item response tree models to investigate acquiescence and extreme response styles in Likert-type rating scales. *Educational and Psychological Measurement*, *79*(5), 911–930. https://doi.org/10.1177/0013164419829855

Paulhus, D. L. (1991). Measurement and Control of Response Bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59) Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

Plieninger, H., & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

Podsakoff, N. P., Spoelma, T. M., Chawla, N., & Gabriel, A. S. (2019). What predicts within-person variance in applied psychology constructs? An empirical examination. *The Journal of Applied Psychology*, *104*(6), 727–754. https://doi.org/10.1037/apl0000374

Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Mehrdimensionaler Befindlichkeitsfragebogen. [Multidimensional mood questionnaire]*. Hogrefe.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819322

Trierweiler, L. I., Eid, M., & Lischetzke, T. (2002). The structure of emotional expressivity: Each emotion counts. *Journal of Personality and Social Psychology*, *82*(6), 1023–1040. https://doi.org/10.1037/0022-3514.82.6.1023

Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470. https://doi.org/10.1177/0963721414550706

van Herk, H., Poortinga, Y. H., & Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346–360. https://doi.org/10.1177/0022022104264126

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

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

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports: Response processes in personality data. *Journal of Personality*, *84*(4), 461–472. https://doi.org/10.1111/jopy.12172